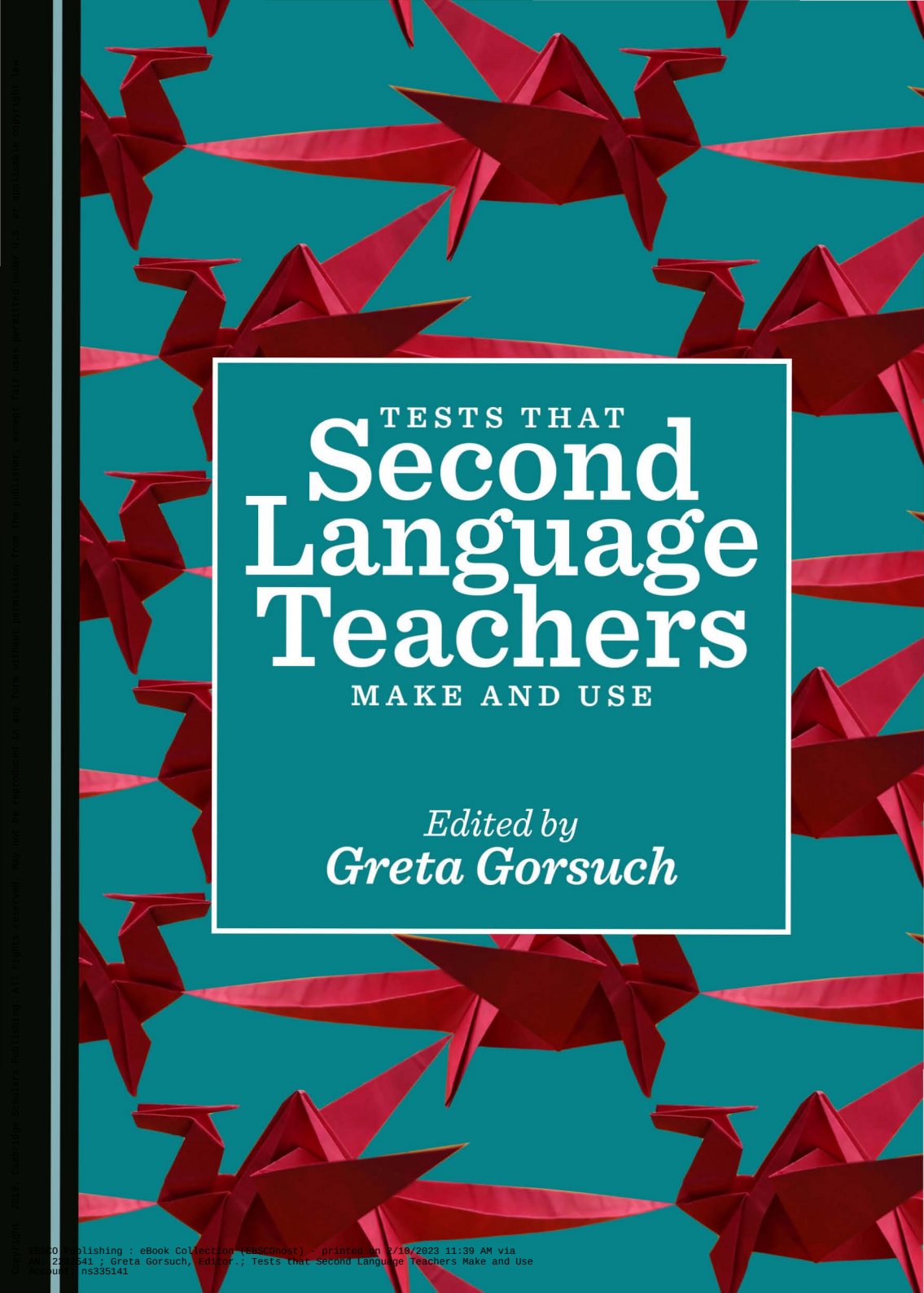


Copyright © 2013. Cambridge University Press. All rights reserved. May not be reproduced in any form without permission from the publisher, except fair uses permitted under U.S. or applicable copyright law.

EBSCO Publishing : eBook Collection (EBSCOhost) - printed on 2/10/2023 11:39 AM via
AN: 2[illegible]541 ; Greta Gorsuch, Editor.; Tests that Second Language Teachers Make and Use
Account: ns335141

Tests that Second Language Teachers Make and Use

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Tests that Second Language Teachers Make and Use

Edited by

Greta Gorsuch

Cambridge Scholars Publishing

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Tests that Second Language Teachers Make and Use

Edited by Greta Gorsuch

This book first published 2019

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Copyright © 2019 by Greta Gorsuch and contributors

All rights for this book reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the copyright owner.

ISBN (10): 1-5275-3901-6
ISBN (13): 978-1-5275-3901-3

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

TABLE OF CONTENTS

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Russian

Spanish

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

CHAPTER ONE

INTRODUCTION: TESTS AND TEACHERS, TESTS AND TEACHING

GRETA GORSUCH

Tests and Teachers

Like it or not, we are "in a relationship" with tests. As learners we have early and continuing experiences with tests. These could have been large-scale standardized tests that all of our school friends took, in all grade levels, in all classrooms, on the same day. These could have been the weekly quizzes we had in high school French, where we had to conjugate verbs. In our lives today, these could be on-site, online tests that we take at the driver's license bureau. Or, our employer requires us to take courses on current laws that are relevant to our roles as employees. We then take a short quiz to certify that we have a basic, as-needed, applied knowledge of those laws.

As second language teachers, our school may require us to prepare learners for large-scale tests so that they may enter university, or graduate. And we have to write and give tests in order to award English-, or French-, or Japanese-language course grades. We may become certified interviewers or graders for a large-scale second language test offered by a testing company.

In some respects, teachers have an uneasy relationship with the simple existence of tests. There are multiple reasons. First, there are those Language Tests, made by commercial testing companies. They command a lot of attention. They are expensive, they may be unfair, teachers may be judged by their learners' results, and preparing for them may consume precious class time. Second language classes are often under-scheduled in relation to what our schools want learners to achieve. So, this last point is particularly painful. Some may feel that such tests capture learner knowledge or skills that the teachers, and even the learners, feel is irrelevant to learners' ability to use a second language to do what they want.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Second, teachers are required to make tests for the purpose of giving grades. Making tests is effortful, time-consuming, and requires technical skills teachers may not feel they have. Making tests and quizzes means that teachers must put into concrete form what they intuitively believe their course is about, and what their learners are about. Such knowledge, called teacher theory by some, is very effective at quickly solving problems, and turning "round" or shapeless content into a linear, moment-by-moment classroom experience. Pinning down our quicksilver teacher theory and turning it into a test, however, is a different matter, and it is challenging to do. And what about the technical skills needed to write and score tests? Are such skills unreachable, and unreasonable to expect working teachers to have? The commercial, large-scale tests may be held up, by ourselves, and by others, as exemplars of "good testing." Yet they are made and scored by professionals, who have specialized training and whose only job is to do testing. This does not often match the reality of teachers' lives.

Third, teachers must balance what they want to test, with what they can test. They may want to test learners' second language conversation ability. They know that their school's required end-of-year test does not capture conversation ability. And yet they also know learners spend a lot of time on conversation practice in classes. But with thirty students, there is no time to give and score such a test. The logistics are impossible. Fourth and finally, giving tests is socially fraught. Teachers must award course grades, presumably based on quizzes, tests, and other sources of information. The resulting scores and grades are bound to create conflicts with learners, and their parents in some cases. Arguing over test scores and course grades may become a major topic of office hour visits. To the extent that our decisions based on tests are called into question, we may have to defend our tests. Even the prospect is deeply unpleasant.

Tests and Teaching

In other respects, second language teachers have a potentially positive relationship with tests. This could be called a "teaching relationship" with tests. There are second language teachers who enjoy making tests, and enjoy what the tests can tell them about their learners. There are many such examples offered by the contributors to this book, found in Chapter Four, "The Tests." Looking at tests through this frame, tests are more supportive of teaching than they may seem at first. Teachers make tests to figure out what learners do and do not know, and what they can and cannot do. Since test scores or teachers' impressions of learners' performances on the test may provide information on this, teachers then have the option of changing

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

their instruction. They might repeat content, such as that unit on paragraph writing, or skip future content, such as that activity on indefinite articles, since the students already seem to know it. Or, teachers may combine content from two or more lessons into a more elaborate project, such as learners planning an all-day tour of their city for foreign visitors. The teacher had learned from quizzes on multiple lessons on constructing descriptions, reporting past events, linking past events together, and solving communication breakdowns within these topics, that learners were ready to attempt something harder.

Other teachers may enjoy giving frequent, ungraded quizzes as a way to focus learners' attention, and to prepare them for graded tests. Quizzes can be interactive and game-like, raising learners' energy levels. There is support for these ideas. For instance, test effect is the idea that taking tests and quizzes helps with memory retrieval, reducing anxiety, and spacing out study, therefore promoting learning processes. Another idea is pedagogically worthwhile tests. This is where teachers make tests that they would also feel comfortable using as classroom materials. These ideas will be described in more detail in Chapter Eight, "Practical Methods for Using Tests for Teaching and Learning."

So, second language teachers are "in a relationship" with tests, and this is unlikely to change. The relationship may take many forms, both negative and positive, as argued above. This book, *Tests that Second Language Teachers Make and Use*, is intended to illustrate and explore the positive aspects of the relationship between tests and teachers, and tests and teaching.

What is this Book?

Welcome to *Tests that Second Language Teachers Make and Use*. This book is a collection of fifteen actual second language tests that working teachers in six countries have made and used. Many of the tests, or parts of them, are still in use. They are quizzes, series of quizzes, unit tests, mid-term exams, and final exams. In other words they are the common classroom tests used by teachers to assess learners' achievement, to provide feedback to learners, and to award course grades. As such, they are criterion-referenced tests and performance tests, that have, in the eyes of the teachers using them, some relationship to the curriculum upon which the course is based. The tests in this book are not used for program-level decisions such as learner placement or admission to a program. They are not used to compare learners in different classes or programs to each other. These functions would be the work of norm-referenced tests, or general language ability or proficiency tests, such as the TOEFL (Test of English as a Foreign

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Language, Educational Testing Service, 2019a), IELTS (International English Language Testing System, IELTs, 2018), or TOCFL (Test of Chinese as a Foreign Language, 2011).

The title of this book uses the term "second language" instead of "second and foreign language." Traditionally, "second language" suggests that a language is being learned in a country where the language is widely used, such as Turkish students learning German in Germany. "Foreign language" suggests that a language is being learned in a country where it does not have widespread or official uses, such as learners of English in Korea. There can be little doubt that teachers and learners in either second or foreign language learning contexts have different experiences and operate under different constraints. However, learning takes place in learners' heads no matter what context they are in. What goes on is grappling with a language other than the mother tongue--a second language. Therefore, the term second language in this book will refer to a second or foreign language, learned in diverse contexts.

The tests in this book are for Chinese (Mandarin), English, French, German, Italian, Japanese, Russian, and Spanish. For non-English language tests, the contributors have provided a basic English translation. The tests reflect the variety of second language courses that teachers teach, and the variety of programs and schools in which they teach. Thus, the tests reflect diverse conceptions of knowledge and skills, content, and learner ability levels.

The tests are coupled with commentary written by the contributors and edited by the book editor. The commentary was built on a principled framework that gets at processes teachers use to write and use tests. See Chapter Two for details on the framework ("A Framework to Probe Tests that Second Language Teachers Make, Score, and Use").

There are five testing content chapters in the book that define key terms and concepts. The chapters are described in more detail below, but they include: "Criterion-referenced Tests and Performance Tests" (Chapter Three); "Communicative Competence and Language Use Description Frameworks and Second Language Tests" (Chapter Five); "Practical Methods for Validating and Improving Tests" (Chapter Six); "Practical Methods for Setting Cut Scores and Making Decisions" (Chapter Seven); and Practical Methods for Using Tests for Teaching and Learning" (Chapter Eight). The chapters present selected, currently and commonly accepted concepts in testing, such as you would find in many available books on language testing. But *Tests that Second Language Teachers Make and Use* goes one step further. The testing concepts in the chapters are illustrated by the contributors' tests and their test commentaries. The commentary probes

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

their processes of test construction, use, and validation. Thus, readers will see testing concepts and practices in action, in the busy, messy, time-sensitive lives of working teachers. One thing the content chapters do not do, is to support some sort of "test police" analysis. The tests here are not held up against some standard implied by the testing content chapters and the framework used to form the test commentaries. The content chapters are offered as help, and illustrations of what working teachers do. The commentary framework aims at consistency and descriptiveness.

This last idea speaks to one of the reasons this book was done. Tests are an everyday feature in the lives of high school and college-level teachers and learners. Nonetheless, teacher-made tests are under-studied. Perhaps the tests are considered "routine" and idiosyncratic. Or, the tests are seen as poor reflections of what testing professionals can do.

Why this Book?

Classroom tests are an everyday feature of second and foreign language high school and college-level classrooms across the globe. Otherwise known as criterion-referenced tests or performance tests, such tests are the familiar quizzes and exams that teachers use. Yet little is known about how teachers make these tests. What are the processes by which they write tests? What knowledge sources do they draw from? If teachers inherit a test from a supervisor or a previous teacher, what do they make of it? Do they use previous tests or parts of them? What is their process of adaptation, and what knowledge sources do they draw from for that?

Still less known are teachers' preferences about test item formats to use, and areas of learners' communicative competence they capture in their test items. Further, little is known about whether, or how, teachers check the reliability (consistency) of their tests, and whether and how teachers validate their tests. These last two speak to the fairness of tests, and the usefulness of the resulting test scores. Finally, instructional staff in high school and college-level foreign or second language departments are diverse in their language backgrounds and levels of professional development. Some instructors have years of experience, whereas others are being supported as novice instructors and graduate students. What can their test making and test use practices tell us about their different understandings of language learning, and program priorities and traditions?

Rather than view teachers' tests as idiosyncratic and poor shadows of what professional test writers can accomplish, it is more constructive to identify the deeper patterns and order of tradition, innovation, and reasoning

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

behind teachers' tests and testing practices. Thus, this book is intended as a resource for readers.

Who is this book for?

This book is intended for multiple audiences, including working teachers, teacher educators, and department-level administrators who want to increase their ability to write good, useful, and defensible classroom tests. Teacher educators and department-level administrators also want ways to use tests as a means of professional development (for themselves and for those working under them), and as measures of course and program success. This last speaks to evaluation, which is specialized, applied research of growing importance to the field of second language education. *Tests that Second Language Teachers Make and Use* may help any of these audiences, in multiple ways.

There certainly are scholars interested in second language teachers and testing (called assessment, by some, to describe less formal "language sampling" activities done by teachers, see Rea-Dickins, 2004). They may find this book useful. It should be stated, however, that the tests and commentaries in this book are not intended to be representative of any sort of pattern or consensus of testing attitudes or practices among teachers. Groups of teachers are not compared in terms of their length of experience, geographic location, learner population, language taught, or previous training, even though they provide information on these points in the test commentaries. There was no intention of establishing a predefined sample of responses, or attempting generalizability with that sample. In other words, this is not survey research. Rather, the editor sent out a call for contributions to every conceivable association, listserv, or website she could find, and to every teaching or teacher preparation colleague she could think of. The collection of tests and commentaries here represent what was sent in response to the call, and what was successfully developed in a collaborative fashion between the contributors and the book editor.

What is in the Book

There are five parts to the book. There are five testing content chapters; a chapter outlining the basis for the test commentaries and the questionnaire used to form them; the tests and commentaries themselves; a reference section; and an index.

The testing content chapters. The testing content chapters have a parallel structure. There is a brief introduction which states the reason for

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

the chapter; definitions of key concepts, which includes examples of applications of the concepts emerging from the tests and commentaries themselves; a chapter summary; and a brief list for further reading.

Chapter Three is on criterion-referenced tests (paper and ink tests) and performance tests (tests where learners write or speak, and their performance judged using scoring criteria). These tests are used to diagnose learners' strengths and weaknesses and give feedback in relation to a course, and to estimate learners' achievement in a course. This chapter comes early, in that criterion-referenced tests and performance tests are the types of tests that appear in this book. This chapter defines concepts needed to understand the function and scope of the teachers'/testers' tests that appear in Chapter Four. It would be less useful to read the tests and commentaries thinking about how the tests would be used for placement in a program, or program admission. Norm-referenced tests are not treated in this book. Because the concepts of criterion-referenced tests and performance tests are foundational to understanding the educational roles and scope of the tests appearing in this book, this is the only testing content chapter that precedes the commentary framework chapter (Chapter Two) and the tests and commentaries (Chapter Four). The remaining test content chapters follow Chapter Four. In this way, readers can digest the tests and commentaries in the order of their own interests, and then consult the remaining testing content chapters as a reference for further understanding.

Chapter Five is on relating test writing to the high- and middle-level theories currently in use in second language education. The high-level theory here is, of course, communicative competence, the idea that language is not only various types of knowledge, but the ability to use a second language. The middle-level theories refer to language-use frameworks such as the ACTFL *Guidelines* (2012a) and *CEFR* (Council of Europe, 2001, 2018). Despite its theoretical impact and professional salience, many working teachers grapple with how communicative competence gets translated into actual test items and tasks. How does textual competence, for example, become test items or a test task? The same may be said of the ACTFL *Guidelines* and *CEFR*. How does a description of what learners are able to do at a particular ability level become test items, a test task, or scoring criteria or rubrics? Perhaps a missing link is understanding better how teachers relate these high-level and middle-level theories and frameworks to their course outcomes. By extension, then, what classroom activities and tasks do teachers then extrapolate from outcomes, and/or the materials available to them? And to what extent then do the activities, tasks, and materials inform the test items or test tasks teachers write? There are probably a host of other missing links that cannot be captured even by

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

complex and nuanced models such as the Teacher Theory Model (Figure 2-1). Some of the test commentaries from contributors to Chapter Four bring further clarity to these.

Chapter Six is on practical methods for validating and improving tests. Validation refers to how teachers/testers know whether a test measures what they think it measures, and whether using scores from a test is appropriate. Validation sometimes brings on images of professional testers hunched in front of their computers, doing statistical analyses beyond what ordinary teachers can do or have time to do. But many validation strategies can be surprisingly practical, and easy to build into tests during the test writing, administration, scoring, and decision-making stages. Validation strategies will be described and illustrated in tests and test commentaries.

Chapter Seven is on practical methods of setting cut scores. Many teachers score their own tests, but in addition they must make decisions based on test scores. In other words, does a learner pass or fail a test? What grade do they get on a test? Practical methods of cut score methods are given adapted to both pass/fail decisions and grade decisions. Examples from tests and commentaries are highlighted.

Finally, Chapter Eight is on practical methods of using tests for teaching and learning. This chapter is essential to taking a more proactive relationship between tests and teaching. In other words, what are ways tests can be used for teaching? Not only to plan instruction, but to use tests as part of teaching? There are exciting developments in general education, and second language education, that contribute to our practical, working understanding of how to more closely marry tests to our courses. Examples of these kinds of thinking will be illustrated using tests and test commentaries from the book.

Probing teacher-made tests, and test and test commentary chapters. These two chapters appear sequentially. Chapter Two, "A Framework to Probe Tests that Second Language Teachers Make, Score, and Use," introduces two models. The first is the Teacher Theory Model (Figure 2-1), first formulated by Griffee (2012a), which proposes explanations of how teachers develop action-oriented theories in schools to solve the "problem" of classroom tests. The second model is a Life of a Classroom Test model (Figure 2-2) which proposes an explanation, ordered sequentially in terms of the life of a test, of how teachers can plan, score, and use test scores to teach. Both models were needed to design the questionnaire that probes the tests the teachers/testers offer in Chapter Four. The models' components, and relationships between them, are defined. Then, the five-part questionnaire is offered.

The second chapter in this sequence, Chapter Four, "The Tests" chapter, is the focal point of the book. Each test appears with a brief introduction

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

naming the teacher/tester, a description of the context in which the test is used, the tests themselves and test answer keys and scoring rubrics, and the teacher/tester's responses to the five-part questionnaire, rewritten as prose under the headings suggested by the questionnaire parts.

It is expected that if any part of the tests published here are adapted or used by readers, the readers will then give credit by listing the proper citation of the work on the test or parts of the test they adapt. Here is how to cite a test from this book using APA (American Psychological Association) Guidelines:

Last name, First name (XXXX). Title of test. In G. Gorsuch (Ed.), *Tests that second language teachers make and use* (pp. xx - xx). Newcastle-upon-Tyne, United Kingdom: Cambridge Scholars Publishing.

References. The reference list offers full citations for previous research or commentary cited in-text. The references are both practical and theoretical, and every effort has been made to list references that are readily available either online, or through a standard library document delivery service.

Index. The index is intended for readers to identify page numbers in the book on key topics, concepts, and definitions they are interested in.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

CHAPTER TWO

A FRAMEWORK TO PROBE TESTS THAT SECOND LANGUAGE TEACHERS MAKE, SCORE, AND USE

GRETA GORSUCH

Why a Framework to Probe Classroom Tests?

This chapter introduces two models. One proposes an explanation of how teachers develop action-oriented theories in schools to solve the "problem" of classroom tests. The second model proposes an explanation, ordered sequentially in terms of the life of a test, of how teachers can plan, score, and (perhaps) use test scores to teach. Both models were needed to design the questionnaire that probes the tests the teachers/testers offer in Chapter Four. The questionnaire appears at the end of this chapter. Beyond this book, models are also needed to organize, understand, and further discuss complex phenomena or concepts in second language education. And teaching, and teaching and testing, are complex.

Explaining how Teachers Develop Theory

The first model aims at illustrating how teachers develop theory. The model is about teachers, and its basis is how teachers learn, and how they create their own personal theories in order to plan and do coherent actions such as teaching, and making tests. See Figure 2-1. Definitions of the components follows.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

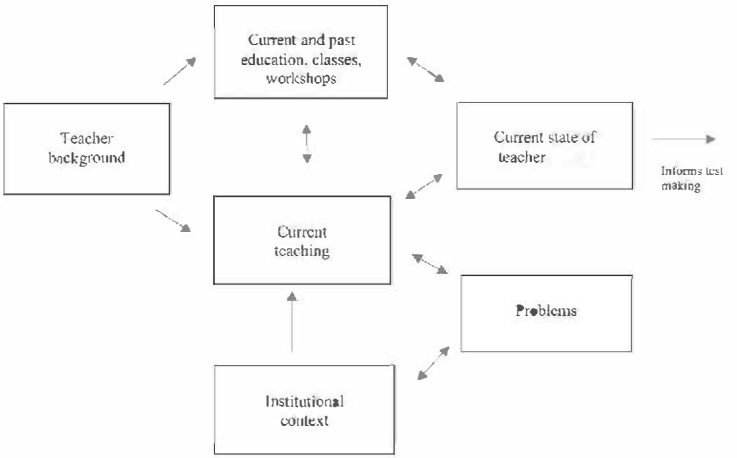


Figure 2-1. **Teacher theory model (based on Griffee, 2012a).**

The teacher theory model originated in research done by Griffee (2012a), who was working with novice second language teachers who were also M.A. in applied linguistics students. His purpose was to learn how novice teachers "construct a teacher-identity that serves as a context for [their] decisions and actions" (p. 202). In related research, Griffee refers to teacher-identity as teacher theory, or low theory, which is "implicit theory held in the minds of individual practitioners" (2012b, p. 52). There will be more detailed discussion of teacher theory, or low-level private theory, and other types of middle- and high-level theories in second language education later in the chapter in the definition section on the model component "Current and past education," and also in Chapter Five on how theories and frameworks of language use descriptions may act as resources for making tests.

Teacher Theory as a Basis for Teacher-made Tests

In Figure 2-1, the component, "Current state of teacher theory" represents teacher theory. Gorsuch and Griffee (2018) refer to low-level theory or teacher theory as "local theory" in that teacher theory is strongly formed by specific contexts with specific learners in courses with a particular curriculum. As can be seen in the model, teacher theory emerges from "Teacher background," "Current and past education," the demands of

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

"Current teaching" and the "Institutional context," and what Griffee identified as "Problems" (2012a). Thus, teacher theory changes over time, depending on current teaching, among other things. Teacher theory is not a new concept in our field. Widdowson (1993, p. 264) refers to a teacher belief system, or "personal constructs of teachers," representing teacher attitudes, thinking, and decision making.

It is argued here that "Current state of teacher theory" forms the basis of teachers making and scoring tests, and using test scores. How is this the case? Traditionally, teacher theory is thought to be about teaching. It is oriented toward solving problems, such as how to understand what learners need and how to deal with it, and what materials to choose and how to teach with the materials (Gorsuch & Griffee, 2018). Teacher theory seems oriented to translating what teachers think a course is about (the course outcomes and course content), into the beginnings, middles, and ends of successive lessons over time. Teachers create activities and tasks that accomplish the lessons. Griffee (2012a) never specified that his teacher theory model represented teacher theory about tests. And Rea-Dickins (2004), who is no stranger to working language teachers and to teacher theory, calls for a model of teachers and tests. This implies that somehow, teaching must be understood differently than testing (or what Rea-Dickins calls "assessment"). Certainly, Nikolov (2016) found that teachers in Hungary, tasked with writing diagnostic tests for young second language learners, produced tests that "tended to focus on errors, accuracy, and what students cannot do" (p. 75). She felt this was different from the *CEFR* frameworks (Common European Framework of Reference; Council of Europe, 2001) that had been adopted at the teachers' schools, which was thought to guide teaching. Thus, in her mind, teachers' teaching and testing were different.

Yet clearly teachers have theories about tests. Barrette (2004); Cheng, Rogers, and Huiqin (2004); Davison (2004); Kikuchi (2005) and others have reported on teachers' tests and testing practices in various countries. These observed teacher-made tests and testing practices would not exist if teachers did not have theories about tests. There are patterns, and the patterns, the tests, and the testing practices are coming from somewhere. Barrette (2004) commented that second language achievement test drafts she studied for her report were influenced "by the course materials and syllabus, the instructor's preferred teaching methodology, [and] the student population" (p. 58). These are all elements of teacher theory (Figure 2-1), and her comments suggests some desired connection between teaching and testing. However, she also cautions that many tests she examined for her report also had fidelity to "teachers' assumptions about

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

what a test should be as opposed to matching test items with what learners do in class" (p. 67). So clearly, teacher theory, which traditionally relates to instruction, also generates teachers' apparent assumptions and decisions and practices and attitudes about tests. It may be the case that the products of teacher theory, whether instruction or tests, are different, depending on teachers' education. Note recent calls for teaching practices, informed by language use description frameworks, to more closely link to classroom tests (Nikolov, 2016), and for more support for teachers to learn testing concepts (Brindley, 1997; Edelenbos & Kubanek-German, 2004; Newfields, 2006). See also Moodie and Haany (2008) for a description of a second language conversation performance test that was overtly fitted to what learners experienced in class, through careful planning.

Definitions for the Teacher Theory Model

In this section, the components of the Teacher Theory Model (Figure 2-1) are defined and related to each other. First, some general comments: The general orientation of the model is from left to right. For example, "Teacher background" is on the far left, appearing to be a foundation for "Current teaching" and "Current and past education." "Current state of teacher theory" appears on the far right, suggesting an end state of some sort, generated by "Teacher background" and mediated by "Current and past education," "Current teaching," and other model components. Single headed arrows suggest that one component influences another, while double headed arrows suggest that two or more components have a potential mutual influence. For instance, it would be hard to imagine that "Current teaching" would influence "Teacher background." "Teacher background" existed before "Current teaching." Thus, there is an implied chronological sequence. But, "Current teaching" might influence what summer training workshop a teacher might choose to take ("Current and past education"). And in a turnaround, "Current and past education" will be drawn upon by a teacher to interpret and guide their "Current teaching." Finally, none of these hypothesized relationships are set in stone, and the components themselves have fuzzy boundaries. For instance, "Institutional context" might be seen as being very similar to "Problems." Is it not within an institution that problems occur, and must be solved? This is what makes models so useful. They create and focus discussion.

Teacher background. The definition for "Teacher background" in Figure 2-1 would be aspects of a teacher's history that are relevant to their current role as a second language teacher/tester, including language learning experiences with the target language (Golombek, 1998; Griffee,

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

2012a), second language ability or status (high competence versus low competence, native speaker versus non-native speaker) (Frain, 2009; Kikuchi, 2005), experiences with tests as a language learner (Barette, 2004), and amount and types of teaching/testing experience (Frain, 2009; Griffee, 2012a; Phillips & Abbott, 2011; Shepard, 2000a). This last aspect, that of teachers' experiences with teaching and testing, is of particular interest because it speaks to "cultures" of teaching and testing that may operate beyond the level of the institution. Cheng, Rogers, & Huiqin (2004, p. 362) noted that second language teachers in Hong Kong and China wrote classroom tests and used "procedures" that "tend to mirror those of the external tests," such as the TOEFL, whereas Phillips and Abbott (2011) noted that American second language teachers had *yet* to be influenced by the ACTFL OPI in terms of their testing practices. The ACTFL OPI (American Council on the Teaching of Foreign Languages Oral Proficiency Interview) is a standardized interview test that has been in widespread use since 1984 (ACTFL, 2018). The two single headed arrows in Figure 2-1 reflect the hypothesis that "Teacher background" influences teachers' "Current and past education" and "Current teaching."

Institutional context. The definition for the component "Institutional context" in Figure 2-1 is related to Golombek's conception of "context" (1998, p. 452) which "includes the institutional and sociopolitical setting along with the time, place, and actors within the setting." This has to do with teachers' knowledge of the institutional setting, such as expectations of teachers, the sorts of learners there are at a school, and what the second language program is supposed to accomplish. What is meant by learners "accomplishing" something in a program may be stated on school websites, etc., but is more likely unstated, and the two might not be the same thing (Griffee & Gorsuch, 2016). Teachers' knowledge of this might be unspoken, but they would know it or have a sense of it after spending time in a program.

Knowledge of an "institutional context" also has to do with constraints on teaching and testing imposed and supported by a specific institution, such as large class sizes, required tests, required materials, etc. Not all institutional contexts impose negative constraints. A school might invest in having smaller classes, and better teacher access to appropriate technology. But often the institutional context is seen as potential clashes "between personal values and institutionalized role requirements and expectations" (Bullough, 1989, p. 79). In other words, the institutional context is the reality in which teachers must operate, and carry out their "Current teaching." If a school wants teachers to use a particular textbook, then a teacher needs to figure out how to use it in line with his or her

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

teacher theory. Such a textbook might (or might not) supply ideas on test content and perhaps test item formats (fill in the blank, short answer, etc.)(Barrette, 2004; Phillips & Abbott, 2011).

The "Institutional context" partly accounts for the context-dependent nature of teacher theory. When teachers proceed in their careers to teach at another school, perhaps in another country with a different set of learners, they experience a period of disorientation as they further develop their teacher theory to learn the new institutional context (Gorsuch, 2007; Gorsuch & Griffee, 2018). The single headed arrow from "Institutional context" to "Current teaching" reflect the hypothesis that the institutional context influences "Current teaching," as given in the examples above. The two headed arrow between "Institutional context" and "Problems" suggests that whatever "Problems" occur (unmotivated or unprepared students, for example, Griffee, 2012a), may be created or mediated by an institutional context. Conversely, the institutional context may also provide an answer to problems teachers encounter, such as having a supervisor or colleague who can help (Griffee, 2012a).

Current teaching. The definition for the model component "Current teaching" is "whatever courses the teacher is teaching" (Griffee, 2012a, p. 217). This invokes teachers' engagement with the curriculum, syllabus, materials, and outcomes of specific courses, and more importantly, their "interactions with students" in those courses (Griffee, 2012a, p. 217). Current teaching is a powerful impetus in "Current teacher theory." It provides an immediate context and need for goal-directed lesson planning, and planning and giving tests and quizzes. Whatever a teacher's current teaching assignment happens to be, this will change what he or she does with quizzes and tests. Teaching courses with large class sizes may constrain a teacher from doing speaking tests with students individually. They may opt instead to record pairs or groups of students talking to each other (Moodie & Haany, 2008; see Grogan and also Shaver in this book). They may rely more on self- and peer-assessments of those recordings, not only due to logistics, but also because they believe self- and peer-assessment contributes to language learning, and is in line with their conception of the course outcomes (Venema, 2002). An instructor with smaller classes may opt to let learners try a speaking task twice, which may result in greater learner attention form and potentially better performance (Hawkes, 2012). One instructor, De Silva (2014), created scoring scales with five performance levels for both writing and speaking performance tests due to the heterogenous, mixed ability level classes she was assigned to teach. She also believed that teaching the scoring scales to

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

learners would support their growth in self-assessment, an apparent course outcome.

"Current teaching" is connected to every other component in the Teacher Theory Model (Figure 2-1). The single headed arrow from "Teacher background" to "Current teaching" suggests that a teacher's background influences what classes a teacher can teach and will seek out to teach. The double headed arrow between "Current and past education" and "Current teaching" suggests that a teacher's past or current education will influence not only what classes they are qualified to teach and can teach (the arrow going in one direction), but also how they teach them and use tests in them (the arrow going in the other direction). Teachers will undoubtedly draw upon classes and workshops they have taken to make day-to-day decisions about their teaching and testing. For instance, Edelenbos et al (2004) claim that German-language teachers in the Netherlands working in classrooms in the 1990s had little testing training in that teacher education policy "had opted against" such training (p. 270). Some teachers in the study said that they used diagnostic testing techniques "embedded" into their instruction (asking learners questions to check understanding, for instance) but the recordings of the classes they taught showed little evidence of it. On a more positive note, Purpura (2016, p. 202) believes that the Common European Framework of Reference for languages (Council of Europe, 2001) has improved the "assessment literacy" of teachers across Europe (see Chapter Five on Communicative Competence and Language Use Description Frameworks and Second Language Tests). The single headed arrow from "Institutional context" to "Current teaching" suggests that it is the institution that determines class size, teaching load, and course assignments for teachers.

Problems. The definition for the model component "Problems" is issues or challenges that arise as a function of "Current teaching." For instance, Griffee (2012a, p. 223) learned that novice teachers' response to learner's lack of motivation and uncommunicativeness in classes was to retreat "to a survival mode teaching vocabulary and grammar." Gorsuch and Griffee (2018) found a more constructive response from an experienced second language teacher who worked with a learner who had particularly challenging pronunciation problems. When the learner failed a high-stakes test, the teacher was confronted by an advisor from the learner's department. This is an example of "Problems." Her response to the advisor was to identify problems in specific areas of the learner's performance on the test. She also pointed out features of the test and testing procedure, such as having two independent raters, that ensured consistency of the learner's score. The double-headed arrow going from

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

"Current teaching" to "Problems" suggests that problems unique to the classes a teacher is working with will arise, and that problems the teacher notices may change their teaching.

Current and past education, and Current state of teacher theory. The two final definitions here are for "Current and past education" and "Current state of teacher theory." They have been touched upon in the previous definitions. Most components of the Teacher Theory Model (Figure 2-1) are hypothetically interconnected (shown as arrows), and thus elements of the definitions are interconnected. But brief, additional definitions are given here. The "Current and past education, classes, workshops" component is any current or past courses, seminars, or workshops a teacher has taken. This could be at the undergraduate or graduate level, leading to a degree or one-off summer workshops. This would also be participation in test development projects that are school-sponsored or sponsored by universities or commercial testing companies, or short training courses intended to familiarize teachers with frameworks of language use description such as the ACTFL *Guidelines* (American Council on the Teaching of Foreign Languages, 2012a) or *CEFR* (Common European Framework of Reference, Council of Europe, 2001, 2018). It will be argued later in this book that language use description frameworks such as the ACTFL *Guidelines* and *CEFR* may constitute a resource for teacher learning in themselves in that they are published and publicized, and construed by some stakeholders in second language education to be a means of establishing accountability (Phillips & Abbott, 2011; Purpura, 2016). Thus, it may not be necessary to have active, externally motivated coursework or workshops in these descriptive frameworks for teachers' theories to be influenced by them.

The "Current state of teacher theory" component represents teachers' action-oriented theories that inform their day-to-day decision-making about testing. "Day-to-day" does not mean that teachers' actions and decisions are random, or lack internal coherence. There are patterns and order, generated by contingent teacher theory, in response to the interaction between "Current teaching," and "Current and past education." The relationships between the three components "Current state of teacher theory," "Current teaching," and "Current and past education" are potentially dynamic, with "Current and past education" about second language testing serving as a possible agent of change with teachers' conceptions of what they do on in the context of their "Current teaching." Yet education is only a potential agent of change. Even with courses and workshops, some testing researchers have found limitations on teachers' "assessment literacy" (teachers' ability to interpret educational literature

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

on testing, write tests, and assess students with "minimal bias" Newfields, 2006, p. 51). Manley (1995) noted that despite the popularity of "oral testing" at conferences and multiple workshops and discussions on oral testing within a school district, teachers had no way to find "an immediate link to the teachers' classrooms" (p. 94). In other words, teachers grappled with how to do oral testing, and what the test scores would mean in terms of the courses they were teaching. For instance, how much weight should be given to a students' grade based on an oral test? Sixteen years later, Phillips and Abbott (2011) reported much the same. Despite the salience of the ACTFL *Guidelines*, teachers had not readily adopted guidelines of achievement into their theory and practice of tests.

The reason education is only a potential agent of change in teacher theory can be explained in part by the complexity represented in the Teacher Theory Model. Rea-Dickins (2004) notes that teachers act in many roles in the classroom and in the institution, and as a result, face "significant dilemmas" in that they are "sometimes torn between their role as facilitator and monitor of language development and that of assessor and judge of language performance as achievement" (p. 253). Here the idea of "achievement" brings us to learners' grades, which teachers may believe requires formal tests and quizzes such as those presented in this book. Achievement also brings to mind teachers' awareness that their school's standing may be construed by administrators as learners' scores on standardized tests. The Teacher Theory Model takes into account the classroom context in "Current teaching" (the need to award grades), as well as "Institutional context" (how administrators construe school standing). Thus, teacher theory is formed by exigencies other than "Current and past education."

Another explanation is how theory is arranged in the second language education field, which is illustrated by Griffee's (2012b) High Middle Low Model (HML Model). He posited three levels of theory: high-level theory, middle-level theory, and low-level theory, each of which is used by different actors in second language education for different purposes. An example of high-level theory would be communicative competence. An example of middle- theory would be more specific domain theories, such as second language acquisition theories, or language use description frameworks such as *CEFR*. Examples of topics addressed in low-level theory (or local teacher theory) are "What works for me and why," or "How my students learn" (Gorsuch & Griffee, 2018, p. 79). See Table 2-1.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 2-1: ***The HML model with examples***

Theory type	Also known as	Characteristics	Examples
High	Grand theories	Articulated in terms most come to agree upon as objective reality, publicly discussed, published Establishes fundamental and overt changes in theory and practice in the field; Answering the question of what reality *should* be	Test validity models and theories Communicative competence Language proficiency models
Middle	Domain theories	Articulated, publicly discussed, published Used to motivate research agendas; May or may not be intended for classroom applications	Theories from communication studies, applied linguistics, education, linguistics, psychology, etc. such as Deliberate Practice Theory, and Exploratory Practice Second language acquisition theories Test constructs Language use description frameworks
Low	Teacher theories	Not articulated, intuitive, not necessarily available for introspection or discussion, private, not published Used to solve the problem of classroom instruction and testing	What works for me and why How my students learn What my students need and how I deal with that What I put on my mid-term exam What I think about students' test scores and do I make future plans according to them

Note. Based on Gorsuch & Griffee (2018, p. 79)

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

It may not be the case that classroom teachers have a lot of experience with high-level and middle-level theories, even in M.A.-level graduate courses. There may have been survey courses with high- and middle-level theories as their content, but such courses may have been taught in generalities, without sufficient support for novice teachers and in-service teachers to apply the theories to the complexities of "Current teaching" contexts (Griffee, 2012a). And it may not be the case that researchers using middle-level theories to motivate their research projects will make specific suggestions about classroom applications for teaching or testing.

It might be argued that the high-level theory of communicative competence has gained a better foothold in teacher's thinking about classrooms. Certainly, communicative competence has been represented in the Common European Framework of Reference over time (Council of Europe, 2001, 2018; see however Fulcher, 2010a). But reports persist of communicative competence being construed in U.S. college departments as just having learners talk to each other about their opinions, without using reading, writing, or listening for communication and personal enrichment (Griffee, 2012a; Swaffar & Arens, 2005). Such is an anemic realization of communicative competence, existing below even the level of conscious notice and discussion (below middle-level theory).

Explaining how Teachers Plan, Write, Administer, and Score Tests, and then Use Test Scores

This second model proposes an explanation, ordered sequentially in terms of the life of a test, of how teachers can plan, write, administer, score tests, and use test scores to teach. As with the Teacher Theory Model, the model was needed to design the questionnaire that probes the tests the teachers/testers offer in Chapter Four. The model is not intended to be prescriptive, but the sequential steps within the model can be used as guidelines. The model is primarily intended to be descriptive. The model takes into account teacher theory (high-, middle-, and low-level theories, Figure 2-1, Table 2-1) that inform test making and use in many contexts in the second language education field, including classrooms, institutions, countries, regions, and transnational companies. See Figure 2-2.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

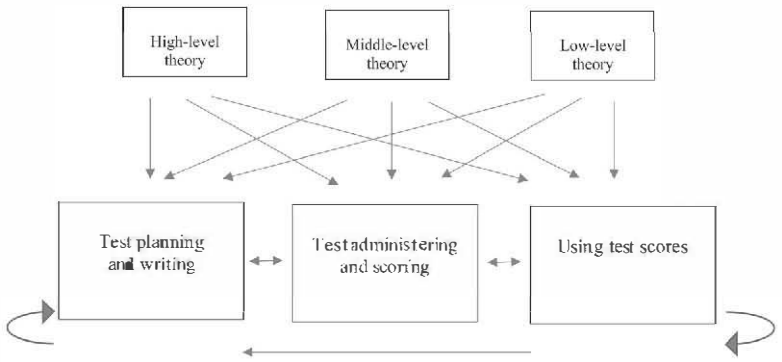


Figure 2-2. **Life of a classroom test model (based on Downing, 2006; Gorsuch & Griffee, 2018; Griffee, 2012a)**

Some general notes: The three larger components on the bottom row represent the focal points of the model, which shows the life sequence of tests. The starting point is the component on the left, "Test planning and writing," which then proceeds to "Test administering and scoring," and then on to "Using test scores." Teachers may then use test scores to return to "Test planning and writing" to improve existing tests, or to use items or test tasks or scoring criteria from the original test to create a new test, hence the arrow circling back to the left and returning to "Test planning and writing." At any point teachers/testers may move leftwards within the sequence. They may begin assembling materials for a test (part of the "Test administering and scoring" component) only to have an insight that would send them back to the drawing board ("Test planning and writing"). Perhaps they found the test directions as written were not clear, or they did not have at hand the right recordings for a listening test. If the teachers/testers had the luxury of giving a pilot test, they may have learned that a test task had seemed clear in the planning stage (listening to a recording and then putting the verbs used into a different form in a different text), but during the pilot learners simply looked at a previous page in the test and got their answers there, as contributor Yesica Amaya found. Within each of the three main components are multiple, more specific, steps that will comprise the definitions for the components detailed below.

The small components in the upper row represent high-level, middle-level and low-level (teacher) theory (Table 2-1). Any of these can inform any stage of testing, hence the multiple arrows leading from each of the

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

theory types to the three focal components of "Test planning and writing," "Test administering and scoring," and "Using test scores." The components of High-level, Middle-level, and Low-level theory are proxies for the Teacher Theory Model (Figure 2-1). In other words, they represent teacher theory. Teacher theory can draw upon high-level, middle-level, and low-level theories ("Current and past education") as well as knowledge of "Current teaching," or information or resources present in the "Institution," etc.

Definitions for the Life of a Classroom Test Model

Within each of the three focal components are specific steps for classroom tests and performance tests that comprise the definitions for the components. The steps were derived from Downing (2006), and from adaptations of Downing's steps set out in Gorsuch and Griffee (2018) specifically for second language teachers.

Test planning and writing. This component is comprised of the steps in which tests are planned and written. Planning includes deciding test content, purpose, item formats, and making plans for establishing consistency of scoring (test reliability). Writing includes writing actual test items for paper and ink classroom tests, and writing, borrowing, or adapting test tasks and scoring criteria for performance tests. The processes of planning and writing may play off of one another in ways that reflect how teachers must balance what learner knowledge or skills they want to capture or know how to capture, against how much class time or personal time they think they have to devote to administering or scoring a test. In other words, they use their teacher theory (Figure 2-1, "Current and past education," "Current teaching," etc.). See Table 2-2.

Table 2-2: ***Test planning and writing for classroom tests and performance tests***

Step	Definition	Examples
1. Overall test plan	The "big questions" stage. This sets the parameters of what a teacher wishes to do, and where questions are answered: What is the test purpose? Who are the test takers? Is the test high stakes or low stakes? What decisions will the test scores be used for? What do I wish to know about learners in	Questionnaire item examples: *Why did you write the test?* *What were the purposes of the test?* *Were you concerned at how long the test would take to administer?* *Were you concerned at how long the test would take to score?*

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

	terms of their abilities? How does the test relate to my course objectives? How does the test relate to what I ask learners to do in the classroom? *For classroom tests:* What test item formats do I wish to use? True/false? Short answer? Do the test item formats "fit" what I want to know about learners? Is there someone I can show my test items to for feedback? *For performance tests*: Is there enough time for me to administer and score the test? Can I get a second scorer? Is there someone to whom I can show my test task and scoring criteria for feedback?	
2. Deciding test content	This is where teachers/testers decide on what their test will sample from the domain of their course. "Domain" in the broadest sense would be descriptions of the knowledge or skills needed to do something in the L2. More narrowly, domain would be communicative functions, texts, tasks, recurring vocabulary and grammar present in a textbook and/or commonly studied or done in class. At this stage teachers/testers might draw from previous tests or language use description frameworks such as *CEFR* (Middle theory) to help define domain with reference to their course outcomes. *For classroom tests*: Includes deciding which chapters or sections from the textbook will be sampled; How a test is related in terms of content and	Questionnaire item examples: *What sources did you draw from for your test items?* Consider: *Test item ideas or content from previous tests? Test item ideas or content from a textbook? Ideas or content from review sections of a textbook? Test items or content from CEFR (*Common European Framework of Reference for Languages*) descriptors? Do you consider learners' communicative competence when writing test items? What aspects of communicative competence? If you wrote a performance test (where learners had to converse or present, or write something above the sentence level), how did you get the ideas for what task learners had to do for the test? From tasks learners do in class?*

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

	activities to the course objectives. *For performance tests*: This may amount to a needs analysis to construct a task that learners must do to produce a performance. Includes what learners need to be able to do, according to course objectives; What learners need to be able to do in the real world.	*From tasks they have do to in real life?*
3. Making test specifications	This is a crucial point in planning, where teachers'/testers' general intentions (Steps 1 and 2) about a test or quiz meet up with specific test items, and performance test tasks and scoring criteria. It is at this stage that teachers/testers state what it is they wish to test (ability, knowledge, skill, etc.) at a level specific enough for others to understand. These statements are then paired with sample test items, or performance test tasks and scoring criteria for the purpose of introspection, and/or getting colleagues' or learners' reactions and feedback. Such test specifications are then used as blueprints for writing more items (Step 4) that will comprise the test. *For classroom tests*: At the item level, a test specification would include a statement such as "learners' ability to hear core vocabulary in spoken statements" and one or two sample items, perhaps in a matching test item format. At the level of the whole test, a teacher/tester would look at the sample items and determine	Questionnaire item examples: *What types of learner knowledge do you believe you are capturing in your test? How does that change with test item types you used on the test? What learner skills do you believe you are capturing in your test? How does that change with test item types you used on the test? Did you seek help from a peer to clarify what your test items or tasks were measuring?* Consider: *Did you state at any point what you wanted to measure in your test? Did you ask another teacher to compare your test items with what you said you wanted the items to measure? Did you make any changes to your test or items as a result of your colleague's feedback? If you wrote a performance test (where learners had to converse or present, or write something above the sentence level), how did you get the ideas for what task learners had to do for the test? From tasks learners do in class? From tasks they have do to in real life? How did you get ideas on how*

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

	whether course core vocabulary is adequately represented in the items. *For performance tests*: A test specification would include a short list of sample tasks that are indicated in a needs analysis, or from a literature review or pedagogical source, or past performance tests (example: a mock employment interview); a list of scoring criteria that could reasonably be used to score learners' performances on a given task (example: ability to respond to questions about past and present employment; ability to give name and address); and a scale (3, 4, or 5 points) to score varying levels of learners' performances. Teachers/testers may choose holistic scales (general and multifaceted descriptions of performance) or analytic scales (specific description of performance), according to course objectives.	*to score learners' performances (the scoring criteria)?*
4. Write test items	Teachers/testers write test items according to whatever steps they took to plan and decide content, instantiating what learner abilities, knowledge, or skills they wish to capture into test item formats or test tasks and scoring criteria. *For classroom tests:* Teachers/testers may write more items than are needed, and then select what they think (or colleagues think) are the best items. *For performance tests*: Teachers/testers choose a test task from a list, according to	Questionnaire item examples: *How did you decide how many test items to write in total?* *Did you write as many items or test tasks as you needed, or did you write more than you needed, then pare down? How did you decide which items or test tasks to keep?* *Did you decide at any point to reduce the number of test items, or test tasks? Why?* *Were you concerned at how long the test would take to administer?* *Were you concerned at how long the test would take to score?*

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

	what task they may think best samples what they most need to know about learners' abilities, and according to what teachers/testers are able to handle in practical terms (they may not have time to arrange for, and score, more than one test task).	
5. Plan scoring	This step is most relevant to performance tests where teachers/testers must balance what they wish to know about learners against what teachers/testers have time to do. For instance, teachers/testers may select only three or four scoring criteria from a list of many possible criteria, due to practical scoring constraints. Will teachers/testers have time to score more than a few criteria while the test is under way? Teachers/testers may also decide at this point whether learners should use the scoring criteria, or some version of the scoring criteria, for self- or peer-assessment purposes.	Questionnaire item examples: *Were you concerned about whether you could get a colleague to help you score learners' task performances? Did you change your plans or your test to reflect time constraints? Were you concerned at how long the test would take to administer? Were you concerned at how long the test would take to score? Did you plan to give the scoring criteria to the learners for future self- or peer-assessment? Did you make another version of the scoring criteria for learners to use?*

While the steps are sequential teachers/testers may move backwards in their planning or writing. For instance, one teacher/tester might plan a performance test to be given three times during the semester, but then, invoking their teacher theory, scale back to giving the test only twice due to time constraints. Another teacher/tester may write matching type items to capture learners' knowledge of core course vocabulary, but after getting feedback from a colleague during the test specification stage, may change to a cloze format, where learners write appropriate answers in blanks in a short reading passage. Perhaps the colleague mentioned that the course textbook was asking learners to read the core vocabulary in short passages.

Test administering and scoring. This component of the model consists of steps teachers/testers take to design and reproduce the documents needed for classroom tests and performance tests and to score

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

the tests. Some of these steps are little-considered because they are seen as mundane. What could feel more like busy work than printing out a test and taking it to the photocopy machine? Yet the steps in this component are consequential. Missteps in composing test directions can cause confusion on the day of the test. Or, teachers/testers may encounter unexpected student responses, in the case of cloze passages mentioned above. What if learners use an unexpected word that is not part of the core course vocabulary and yet is still appropriate for the passage? Finally, learners may question their scores. Are teachers/testers prepared to plausibly defend the consistency and fairness of their scores?

As with Table 2-2, the steps in Table 2-3 are sequential, yet teachers/testers may move backwards, particularly for step 9 ("Administer test") which may comprise a pilot test. A pilot test is something not all teachers have time to do. But when a pilot can be done, the information coming from such an experience will be valuable. For instance, a teacher/tester may find that performance test scoring criteria sheets need a different design, perhaps with more space for raters to write comments. Or they may find they cannot reasonably rate learners on eight criteria in real-time conditions given the short time in which learners need to complete a spoken task. This may send the teacher/tester back to step 6 ("Design and assemble test") or even back to step 5 ("Plan scoring," Table 2-2), resulting in longer tasks or longer time to complete a task, or fewer criteria for scoring.

Table 2-3: ***Test administering and scoring for classroom tests and performance tests***

Step	Definition	Examples
6. Design and assemble test	This step what Downing (2006) calls quality control. The layout of a test should not cause distractions as learners take the test. This is also a time to decide whether to make one or two versions of the test. This could be for security purposes (one group of learners gets version A and another group gets version B), or to deal with learners who do not come on the scheduled day of the test. This is also a time for peer review of a test, to	Questionnaire item examples: *How did you decide how many subtests to write? Was there some correspondence between a subtest and a course objective, or was there some other reason to create the subtests you did?* *How did you decide how many items to write for each subtest? Did you have one version of your test, or did you create a second equivalent version? What did you do to ensure test security?*

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

	ensure that learners are not confused about what they need to do to respond to test items. Colleagues may also have insights as to whether two test forms (A and B) are equivalent in difficulty and sampling the same test content. *For classroom tests*: There should be a final check on whether test item types match what learners are accustomed to doing in class. Or, at least to ensure learners have sufficient practice answering the item types being used. Teachers/testers should consider having separate subtests for each course objective and ensuring these are clearly differentiated on the test paper. Other considerations: Is the artwork clear? Are the instructions on the test itself or on a separate sheet, and can learners navigate them? Does the audio match what is needed on a listening test or subtest? If learners are taking a computer administered test, is the available technology easy to use, and in good order? Are learners versed in taking such tests, and in navigating the computer and test interface? *For performance tests*: Teachers/testers should think through whether learners have had sufficient practice doing the tasks required in the test. Other considerations: Is the technology needed for recording learners' written or spoken responses adequate, and in good working order?	*How did you deal with learners who missed the test, or who were late for the test?* *Did you ask another teacher to look over your test before you administered it? What did he or she say? Did you make any changes to your test as a result?* *Were there any test items or test procedures that were new to learners?* *How did you prepare learners to take the test?*

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

7. Reproduce test	This step is also about quality control. Once the test or performance test scoring criteria sheets are photocopied, are the pages in the right order? Is the print dark enough? Are there enough copies? Is the test secure? Is there any possibility that through the photocopying process, some students may have seen the test or heard about test items? On the day of the test: If test items, or performance test tasks are shown to learners electronically or on the board at the front of the class, can all learners see them?	Questionnaire item examples: *Were you ever concerned about test security? What did you do to ensure test security? Were your tests photocopied, or did learners see your test items or test tasks another way, such as on the blackboard?*
8. Administer test	This step is concerned with anything teachers/testers did to ensure all learners took the test under the same conditions. If a teacher/tester has the time to do a pilot test (a rehearsal test) with learners similar to the intended test candidates, it would be at this stage. Teachers/testers can keep notes on any changes in the test or test administration they made as a result. These changes can be based on learners' comments, or the teacher's/tester's observations.	Questionnaire item examples: *What did you do to ensure all learners had the same conditions under which they took the test?* *Did you later change your scoring because you recognized some learners couldn't answer certain test items due to some aspect of the test?* *Did you pilot your test? Do a trial run? Did the pilot result in any changes to the final version of your test?*
9. Rater training (for performance tests)	This step is to promote consistency in scoring subjectively scored tests. If a teacher/tester has a colleague who can also score learners' written or spoken performances, then the colleague will need training. This may consist of scoring together a sample of three to	Questionnaire item examples: *For performance tests, did you score learners' written or spoken performances with a colleague? Did you compare scores? Or were you the only scorer? Did you score the tests, then set the test asides, and then score the tests a second time?*

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

	five learners' performances on a previous test if the test task, if scoring criteria are the same as the current test. Or, learners' performances from the current test, if there is time between the test administration and reporting test scores. The two raters together should learn the test task and the scoring criteria, then score sample performances, and then compare their scores after the fact. If there are differences, then the raters need to discuss how they arrived at their scores until they reach consensus. This may result in better rater learning of the test task and scoring criteria, and it may result in better-written test tasks and/or scoring criteria. Teachers/testers working alone might score learners' performances, then set the tests aside for 72 hours, and then score the performances again without looking at the first score given.	*Did you make any changes to your scoring criteria as a result of colleagues' feedback during rater training?*
10. Score test	This step is to boost accuracy in scoring, and in the case of performance tests or classroom test items that are subjectively scored, reducing teacher/tester bias. *For classroom tests*: Downing (2006, p. 17) suggests a two-stage scoring process where tests go through preliminary scoring, then teachers/testers make a key based on insights from preliminary scoring, and then a final scoring takes place. The middle key-making stage is when teachers/testers consider alternate answers to	Questionnaire items examples: *For classroom tests: How did you accomplish scoring learners' tests?* *Consider: Did you score learners' tests twice for accuracy?* *Did a colleague help you score?* *Did you write a test key?* *Did you consider alternate answers and add them to the key?* *Did you hide learners' names as you scored?* *Did you go back and change your marks on previously*

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

	items where learners complete or write in responses, such as cloze or short answer. Learners may have written answers, such as vocabulary items, that were unexpected yet are still appropriate. The scoring key would include unexpected answers, gotten from the preliminary scoring, that the teacher/tester decided would be acceptable, perhaps through consultation with colleagues. Answer keys would improve accuracy and reduce grader bias by allowing a colleague to participate in the scoring process, using the key to score the test. This would comprise a peer check. Teachers/testers may also hide learners' names while scoring, or randomize learners' tests into a pile for scoring. Teachers/testers may put learners' responses into a spreadsheet for further specialized analysis (step 11), and in the process note inconsistencies in scoring objective or subjective test items. *For performance tests*: To reduce rater bias, teachers/testers might put in action any rater training they did with colleagues (step 9). Two raters can rate learners' spoken or written performances and then compare their ratings informally for consistency. For spoken performances, teachers/testers may decide to score learners in real-time conditions, but then also audio-record learners to score	*scored tests in response to problems you found while scoring tests later in the process?* *For performance tests: How did you accomplish scoring learners' performances? Consider: Did you record learners' spoken performances to score later, or perhaps score a second time?* *Did you score learners' spoken performances at the same time learners gave their performances? If so, did you have enough time to score? Did you have breaks between learner performances? How did you deal with fatigue? Did a colleague score with you?* *For writing tests, did you ask a colleague to score a few of the tests independently as a way to spot-check your scores?*

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

	again at a later time. Teachers/testers who do not have time for this may build in design features to reduce fatigue, such as making sure they have sufficient scoring time, and building in breaks between "groups" of two or three performances. For written tests, teachers/testers may ask a colleague to score a few of the tests independently as a way to spot-check to control bias.	
11. Estimate reliability	This step may be seen as a technical- and specialist-oriented extension of step 10. Reliability is whether a test can be scored consistently, and thus whether a learners' score can be taken as a true reflection of their abilities, knowledge, or skill. Some of the analyses used by teachers/testers might be statistical and thus beyond the scope of this book (see Gorsuch & Griffee, 2018 for specific procedures). Some of these procedures include item analysis and Cronbach's alpha for classroom tests, and calculating Pearson's Product Moment correlation on two sets of raters' scores on performance tests. Some teachers/testers do these analyses. Other strategies used for steps 9 and 10 such as taking steps to ensure learners all had the same test conditions and consistent scoring procedures may be more commonly used, and will boost test score reliability, and test fairness.	Questionnaire item examples: *How do you know your test is reliable? What did you do to check?* *What did you do to ensure all learners had the same conditions under which they took the test?* *For either classroom tests or performance tests, did you ask another teacher to grade your test? Did you compare the grade you gave, with the grade your colleague gave?* *For classroom tests, did you use item analyses such as item facility, or B-index?* *For a classroom test, did you use any statistical reliability estimates, such as Cronbach's alpha?* *Did you use any statistical reliability estimates, such as Pearson Product Moment to compare scorers' scores on a performance test?*

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Using test scores. This final component of the model represents the steps comprising the sequential end of the life of a test. There is on one hand the potentially uncomfortable institutional aspect of tests, which has to do with accountability: Who passes and who fails? On the other hand, this component speaks directly to classroom practice. The component is not really an ending in this sense. "Using test scores" points to some of the most generative and positive activities connected to testing, and speaks to why we even do classroom testing, completely aside from the institutional duty of awarding course grades.

Using test scores is an area of perennial interest to teachers/testers. Teachers/testers must decide what constitutes passing or failing scores on a test, or what score levels should reflect a grade of A, B, etc. (step 12)(see Chapter Seven). Teachers/testers must report scores, which may become a focal point for defending the fairness and trustworthiness of a test (step 13). Then teachers/testers need to decide how to use learners' scores in the classroom, as potential areas of learning (step 14)(see Chapter Eight). As noted in the introduction to this book, many teachers/testers enjoy what the tests can tell them about their learners and may change their instruction based on learners' scores. The test results may form the basis for feedback negotiated between teachers/testers and learners, which scholars in and out of the field of language education have argued can be a seat of learning (Fulcher, 2010a; Purpura, 2016; Shepard, 2005). See Table 2-4.

Table 2-4: ***Using test scores***

Step	Definition	Examples
12. Establish cut scores	Cut scores on a test divide groups of learners in terms of their status, such as pass/fail; learners who know the material/learners who need more work; and A, B, C, D, or F-grade learners. Testing specialists have practical suggestions for setting cut scores, which may be used by teachers. One example is using consensus between colleagues on what constitutes the performance, in terms of classroom test items and performance test task and criteria, of a minimally competent (C minus) learner. Thus, cut scores are made in reference to the test	Questionnaire item examples: *How did you decide which scores were passing or failing scores (cut scores)? How did you decide which scores meant a specific grade?* Consider: *Did a language use framework such as CEFR or other standards help you determine cut scores? Did your institution stipulate cut scores?* For a performance test, *did you use your scoring criteria to determine cut scores? Did you consult a testing book or think of previous*

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

	itself and the course objectives, meaning that learners are compared to the test. This is different from a perhaps prevailing norm-referenced type sensibility which results in "A" learners getting scores between 90 and 100, "B" learners getting between 80 and 89, etc. Language use description frameworks such as *CEFR* offer descriptions of what learners can do in a second language that may focus collegial discussion and decision-making on cut scores (see Fulcher, 2010a for cautions on using *CEFR* for local standard setting).	*coursework you had to determine cut scores? Did a colleague or supervisor suggest cut scores?*
13. Report scores	This step represents an institutional duty of many teachers/testers to report test scores to learners. Downing (2006) emphasizes that scores should be reported in timely fashion, and that test score interpretation should be done in a clear way that boosts the trustworthiness of a test. Thus, this step may include teaching learners to interpret the test scores. This step also includes teachers/testers interpreting the test scores in terms of learners' course grades, and thus relating a test to the course objectives. Good test planning and writing at the early stages in the sequential life of a test (Step 1. "Overall test plan," Step 2. "Deciding test content," Step 3. "Making test specifications," Table 2-2) greatly helps with using tests to determine course grades, which teachers/testers may find fraught and difficult to do for both institutional and social reasons. This speaks to the summative use of test scores, where the scores are used for accountability purposes.	Questionnaire item examples: *Did you hand the test back to learners? Did the learners get to keep the tests? Or did you take the tests back?* *How did you report scores to learners? Was timeliness of concern to you?* *Did you teach learners to interpret their scores?* *In the case of performance tests, did you use the test criteria to help learners interpret their scores?* *What was the role of the test score in determining learners' grades?* Consider: *How much weight did you give your test? How did you decide?* *How did you use learners' scores from this test?* Consider: *Were the scores for your use only?* *Did anyone else use the test scores? What for?*

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

	Learners' scores may be reported to other stakeholders, including administrators or parents.	
14. Use results/ scores	This final step in the sequential life of a test may represent opportunities for learning, not only for learners, but also for teachers who 1. Want to know what their learners can and cannot do, 2. Want to know if a test changed what or how learners studied, or 3. Want to know how a test, or part of a test, can be improved the next time it is used. For learners, feedback from tests are useful for self-directed learning (one expression of washback). To further investigate washback, some teachers/testers may use test scores for insights on whether learners changed what or how they studied. Teachers/testers may use test scores and the tests themselves to turn into classroom learning activities for review, and for memory retrieval. Teachers/testers may also change their instruction due to learners' test scores, to address harder-than-thought content or skills, or to skip future content because learners already seem to know it. Teachers/testers find specific ideas for revising a test for future use to make it more reflective of a course, or more reflective of the learner population's characteristics.	Questionnaire item examples: *Do you feel you learned what you needed to learn from the test results about what learners could and could not do, or what they knew or did not know?* *Did you spend time going over the test in class?* Consider: *Did you mention trouble points as general comments?* *Did you go over each item?* *Did you go over specific subtests?* *Did learners' test scores change your teaching?* Consider: *Did you re-teach content because learners didn't do well on your test?* *Did you skip content because learners did well on your test?* *Did you re-order content?* *If you could turn back time, what would you change about your test? What would you change about your test administration?*

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

The steps in Table 2-4, to extent that teachers/testers consider them or do them, may reveal insights on how to revise a test for re-use at a future time. Note the arrow looping back from "Using test scores" to "Test planning and writing" (Figure 2-2). For instance, a teacher/tester may decide that with a performance test they used, they wished they had taught learners to more readily understand the scoring criteria. In post-test review sessions, the teacher/tester may have realized learners weren't clear what the scoring criteria meant. Thus, the teacher/tester may consider rewriting scoring criteria for easier interpretation by learners, or teaching the criteria thoroughly in classes preceding the next performance test (see De Silva, 2014 for sample scoring criteria specifically used for learner instruction). Even if teachers/testers do not re-use specific test items or scoring criteria, they will remember their thinking and practices in these steps, which may inform their future test practices, adding to their teacher theory.

The Questionnaire

The questionnaire used to explore contributors' thinking about their tests was developed over a five-month period. The Call for Contributions cast a wide net, allowing for any type of high school- or college-level classroom test in any language. Thus, the questionnaire needed to encompass a wide variety of classroom tests or performance tests. At the same time, the questionnaire items needed to account for test planning, writing, scoring, and score use processes. The resulting questionnaire was long, but it was never intended that teachers/testers complete all items. And while the questionnaire represents many good testing practices suggested in the literature (e.g., Downing, 2006; Gorsuch & Griffee, 2018), it is not intended to be prescriptive. Thus, as can be seen in the questionnaire directions, it is not wrong to leave items unanswered if the items pose inapplicable or impractical practices. Note that the first three major sections of the questionnaire mirror the three major components of the Life of a classroom test model (Figure 2-2), "Testing planning and writing," "Test administering and scoring," and "Using test scores" (see also Tables 2-2, 2-3, 2-4).

Tests that Second Language Teachers Make and Use Questionnaire V3
by Greta Gorsuch

Your name:

Your affiliation:

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Introduction

Classroom tests are an everyday feature of second and foreign language high school and college-level classrooms across the globe. Such tests are the familiar quizzes, mid-term exams, and final exams that teachers use to assess student learning, to provide feedback, and to award course grades. Yet little is known about how teachers make these tests.

This questionnaire is designed to transparently describe the processes teachers use to write, administer, score, and use tests. To create the items for this questionnaire, I consulted multiple sources. These included Barrette (2004), Borg (1999), Brindley (1997), De Silva (2014) Downing (2006), Gorsuch & Griffee (2018), Griffee (2012a, 2012b), Griffee & Gorsuch (2016), Kunnan (1998), Purpura (2016), and Shepard (2000a, 2005).

Instructions

1. Please use this Microsoft Word document to answer the questions. Just type your responses directly onto the document. I will take care of formatting later.
2. Be sure to type your name at the top of the document. If you wish to have your affiliation appear in *Tests that Second Language Teachers Make and Use*, please include it.
3. Complete as many questionnaire items as you can. If an item does not apply, simply type in "Does not apply." If you cannot answer an item because you did not do a particular practice in connection to your test (because you are unfamiliar with a testing practice, or do not see the practice as practical or important or applicable, etc.), simply leave the item blank. There is nothing wrong with doing so.
4. Be as complete with your answers as you can. Here is the timeline for submitting your answers. As you can see, we will be going back and forth on your responses over time. I may ask for clarifications on your responses:

   November 1, 2018: Notification of acceptance and request for written responses to questionnaire

   January 30, 2019: Final version of test and written responses to questionnaire Due

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

February 25 - June 25, 2019: Editing of chapters, negotiation of final chapter prose, based on questionnaire responses

June 30, 2019: Notification of final acceptance

5. Please complete all sections of the questionnaire. There are four sections to the questionnaire: 1. Test planning and writing; 2. Test administering and scoring; 3. Using test scores; and 4. Evaluating and reviewing your answers.

The questionnaire may seem long, and daunting! Please do your best with it. Even if some questionnaire items seem redundant, please answer where applicable. We will develop your responses together, over time. Please note that none of the questionnaire items are intended to be prescriptive. There is no implied "standard." The questionnaire aims at breadth and descriptiveness.

Note: Test "item" means a single question on a classroom test. Test "task" means what learners are asked to do on a performance test.

Contact: greta.gorsuch@ttu.edu with any questions or concerns.

Questionnaire

Section One: Test Planning and Writing

Why did you write the test? What were the purposes of the test?
Consider: To help learners focus their thoughts, or change how they studied?
To award a course grade, or part of a grade?
To learn whether learners met a course objective?
To use the scores to give learners feedback on their progress?
Something else?

How did you decide how many subtests to write? Was there some correspondence between a subtest and a course objective, or was there some other reason to create the subtests you did?

Did you write as many items or test tasks as you needed, or did you write more than you needed, then pare the number down? How did you decide which items or test tasks to keep? How did you decide which items or test tasks to discard?

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

How did you decide how many items to write for each subtest?

How did you decide how many test items to write in total?

Did you have one version of your test, or did you create a second equivalent version?

Were you concerned at how long the test would take to administer?

Were you concerned at how long the test would take to score?

Were you concerned how you might use the test items themselves for learner feedback?

Did you consider having your students take your test on a computer? Why or why not? If not, what were your concerns?
Did you consider making the test an open book test?

Did you consider allowing learners to use additional sources such as dictionaries, or their notes, while taking the test?

Did you plan to allow learners to re-take a test for improvement? The same test, or a different test?

What sources did you draw from for your test items?
Consider: Test item ideas or content from previous tests?
Test item ideas or content from a textbook?
Ideas or content from review sections of a textbook?
Test item ideas or content based on textbook activities?
Test item ideas or content from course objectives?
Test item ideas or content from what learners do in class?
Test item ideas or content from what learners do for homework?
Test item ideas or content from worksheets you make?
Test item ideas or content from computer programs learners use?
Test item ideas or content from hybrid course materials?
Test item ideas or content from work you have done for testing companies, or test committees at your school?
Test items or content from the ACTFL (American Council on the Teaching of Foreign Languages) guidelines or descriptors?
Test items or content from *CEFR* (Common European Framework of Reference for Languages) descriptors?

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you consider learners' communicative competence when writing test items?
What aspects of communicative competence?
Other sources not named here?

Which test item formats do you prefer to use?
Consider: Short answer, fill-in-the-blank, matching, cloze, performance test, etc.?

What types of learner knowledge do you believe you are capturing in your test? How does that change with test item types you used on the test?

What learner skills do you believe you are capturing in your test? How does that change with test item types you used on the test?

If you wrote a performance test (where learners have to converse or present a topic, or write something above the sentence level), how did you get the ideas for what task learners had to do for the test?
Consider: From tasks learners do in class?
From tasks they have do to in real life?

How did you get ideas on how to score learners' performances (the scoring criteria)?

If you used points on a scale for scoring learners' performances, how did you decide on how many points on a scale to use?

Were you concerned about whether you could get a colleague to help you score learners' task performances?
Consider: Did you have ideas about how to ensure score consistency, such as having two scorers (you and a colleague), or scoring learners' performances by yourself on two different occasions?
Did you change your plans or your test to reflect time constraints in terms of test scoring?

Did you plan to give the scoring criteria to the learners for future self- or peer-assessment? Did you make another, perhaps simpler or shorter, version of the scoring criteria for learners to use?

Did you seek help from a peer to clarify what your test items or tasks were measuring?

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Consider: Did you state at any point what you wanted to measure in your test?

Did you ask another teacher to compare your test items with what you said you wanted to measure?

Did you make any changes to your test or items as a result of your colleague's feedback?

Did you compare your test to the lessons that learners had?

Did you compare your test to the textbook or other materials learners used?

Adapting existing tests

Are you required to use specific tests in your program?

Was your test different from required tests? How?

Did you inherit your test or parts of your test?

When you first inherited the test, what did you make of it? The test content? The test items? The test item types?

Did you use items or ideas or content from previous tests?

How did you change the parts you decided to keep? Why?

What did you add that was new, or different? Why?

Section Two: Test Administering and Scoring

Were you concerned about test security? What did you do to ensure test security?

Were your tests photocopied, or did learners see your test items or test tasks another way, such as on the blackboard?

How did you deal with learners who missed the test, or who were late for the test?

How did you prepare learners to take the test? Were there any test items or test procedures that were new to them?

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you pilot your test? Do a trial run? Did the pilot result in any changes to the final version of your test?

Did you write any of the test in the learners' first language? Why?

Was your test administered on a computer?
Consider: What was the advantage of doing so?
Did learners respond on computers?
Was using a computer a common classroom experience for learners?
Was this a new experience for learners? How did you prepare them?
Did using computers for test-taking change your test, or your test administration in any way?
Were there any problems with the technology while learners took the test? What did you do?

For classroom tests: How did you accomplish scoring learners' tests?
Consider: Did you score learners' tests twice for accuracy?
Did a colleague help you score?
Did you write a test key?
Did you consider alternate answers and add them to the key?
Did you hide learners' names as you scored?
Did you randomize learners' tests for scoring, thus erasing any order from where learners sat, or at what point they handed their test in?
Did you go back and change your marks on previously scored tests in response to problems you found while scoring tests later in the process?
Did you later change your scoring because you recognized some learners couldn't answer certain test items due to some aspect of the items or the test?

Did you put learners' responses to items into a spreadsheet for further analysis? Did that process help you catch scoring accuracy problems? Or problems with bias?

Did you ask the students themselves to score their own test? Or a classmate's test?

For performance tests: How did you accomplish scoring learners' performances?
Consider: Did you record learners' spoken performances to score later, or perhaps score a second time? Was recording equipment available?

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you score learners' spoken performances at the same time learners gave their performances? If so, did you have enough time to score? Did you have breaks between learner performances? How did you deal with fatigue?

Did you go back over your scoring a second time, after the test was over?

Did a colleague score with you? In real time, or after the test?

For writing tests, did you ask a colleague to score a few of the tests independently as a way to spot-check your scores?

Do you think your test was reliable? What did you do to check?

Consider: Did you ask another teacher to look over your test before you administered it? What did he or she say? Did you make any changes to your test as a result?

What did you do to ensure all learners had the same conditions under which they took the test?

For a classroom test, did you use item analyses on a spreadsheet such as item facility, or B-index?

For a classroom test, did you use any statistical reliability estimates, such as Cronbach's alpha?

For performance tests, did you do any rater training?

Did you make any changes to your scoring criteria as a result of colleagues' feedback during rater training?

Did you score learners' written or spoken performances with a colleague? Did you compare scores? Or, were you the only scorer? Did you score the tests, then set the test aside, and then score the tests a second time?

Did you use any statistical reliability estimates, such as Pearson Product Moment to compare raters' scores on a performance test?

Reporting scores

Did you report the scores to learners? If so, how did you report test scores to learners?

What was your goal in reporting the test scores to learners?

Did you teach learners how to interpret their test scores?

Did you report the scores to anyone else?

Did you spend time explaining scores, or answering learners' questions about scores in or out of class?

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you report peer-assessment scores or self-assessment scores on the test?

How quickly did you report scores to learners? Was speed a priority?

Section Three: Using Test Scores

Cut scores

How did you decide which scores were passing or failing scores (cut scores)? How did you decide which scores meant a specific grade on a test?

Consider: Did a language use framework such as *CEFR* or other standards help you determine cut scores?

Did your institution stipulate cut scores?

For a performance test, did you use your scoring criteria to determine cut scores?

Did you consult a testing book or think of previous coursework you had to determine cut scores?

Did a colleague or supervisor suggest cut scores?

How did you use learners' scores from this test?

Consider: Were the scores for your use only?

Did anyone else use the test scores? What for? Do you believe this use was appropriate?

Did learners' test scores have any positive or negative consequences for you, in terms of your institution?

What was the role of the test score in determining learners' grades?

Consider: How much weight did you give your test? How did you decide?

Were other measures used to decide learners' grades, besides your test?

What was the relationship of the other measures to your test?

Did your test capture some knowledge, skill, or ability the other measures did not capture?

Reporting scores

How did you report scores to learners? Was timeliness of concern to you?

Did you hand the test back to learners? Did the learners get to keep the tests? Or did you take the tests back?

Did you offer feedback to individual learners in addition to their test scores? Written? Orally? In or out of class?

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you teach learners to interpret their scores?

For performance tests, did you use the test criteria to help learners interpret their scores?

Using test scores

Do you feel you learned what you needed to learn from the test results about what learners could and could not do, or what they knew or did not know?

Did your test change how learners studied?
Consider: Did you use learners' scores to find out if your test caused washback?

Did you use particular item types or a performance test to change learners' practices or support their learning? Did their scores indicate they had changed their learning practices?

Did you spend time going over the test in class?
Consider: Did you mention trouble points as general comments?
Did you go over each item?
Did you go over specific subtests?
Did learners offer answers?
Was going over the test a classroom activity?

Did learners ask you about the test itself (not the test scores) outside of class? If so, what did they want to talk to you about? Did you use specific subtests or items or tasks to focus your talks with learners?

Did learners' test scores change your teaching?
Consider: Did you re-teach content because learners didn't do well on your test?
Did you skip content because learners did well on your test?
Did you re-order content?
Did you change the amount of class time or homework spent on specific content?
Did you change your teaching for future courses based on test results?

If you could turn back time, what would you change about your test? What would you change about your test administration?
Consider: Did learners give you feedback on the test? Did they think the test was fair, or helpful?

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did others (parents or administrators or colleagues) give you feedback on the test?

If you used the test and test scores for additional learning opportunities, did anything about that process help you revise the test for future use?

Section Four: Evaluating and Reviewing your Answers

Please read through your answers to the items and answer the following:

To what extent do you think you've described recurrent patterns in your work with tests?

To what extent is your test here an innovation, or something new, for you?

Chapter Summary

This chapter introduced two models that probe teacher theory (Figure 2-1) and teacher test planning and writing (Figure 2-2). In order to understand how teachers solve the "problem" of classroom tests, we have to understand how they develop in-context, action-oriented theories to begin with, and how the theories may be related to making and using tests. Both models informed design of the questionnaire that probes the tests the contributors offer in Chapter Four. Contributors' responses to the questionnaire items may offer further insights on how teacher theory is linked to testing practices.

Further Reading

In addition to descriptions of teacher theory from Golombek (1998) and Gorsuch and Griffee (2018), the following are about teacher theory. Earlier work on pedagogical content knowledge from general education likely informed some of these later works from the field of second language education.

On Teacher Theory

Elbaz, F. (1981). The teacher's practical knowledge: Report of a case study. *Curriculum Inquiry, 11*, 43-71.

Freeman, D. (1998). *Doing teacher research: From inquiry to understanding*. Boston: Heinle & Heinle.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Gorsuch, G. & Griffee, D.T. (2018). *Second language testing for student evaluation and classroom research.* Charlotte, NC: Information Age Publishing.

Shulman, L. (1986). Those who understand: Knowledge growth in teaching. *Educational Research, 15*(2), 4-14.

Widdowson, H. (1993). Imovation in teacher development. *Annual Review of Applied Linguistics, 13*, 260-275.

Woods, D. (1996). *Teacher cognition in language teaching: Beliefs, decision-making and classroom practice.* Cambridge: Cambridge University Press.

On Teachers Writing Tests

Barrette, C. (2004). An analysis of foreign language achievement test drafts. *Foreign Language Annals, 37*(1), 58-70.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

CHAPTER THREE

CRITERION-REFERENCED TESTS AND PERFORMANCE TESTS

GRETA GORSUCH

What this Chapter is About

This chapter is about criterion-referenced tests (CRTs) and performance tests. Both types of tests are commonly used as classroom tests, even though teachers may not refer to them as "CRTs" or "performance tests." CRTs and performance tests are used to diagnose learners' strengths and weaknesses and to give feedback in relation to a course, which teachers/testers and learners can both act on. CRTs and performance tests are also used to estimate learners' achievement in a course. Scores on CRTs and performance tests, along with other measures, would then be used to award course grades. For this chapter, CRTs are typically paper and ink tests, such as mid-term exams, end-of-chapter tests, and quizzes. Performance tests are tests where learners write or speak, and their performance judged using scoring criteria. Because the concepts behind criterion-referenced tests and performance tests are foundational to understanding the educational roles and scope of the tests appearing in this book, this is the only testing content chapter that precedes the commentary framework chapter (Chapter Two) and the tests and contributor's commentaries (Chapter Four).

This chapter defines four key concepts: Decisions made with tests, norm-referenced tests (NRTs), criterion-referenced tests (CRTs), and performance tests. Examples of the concepts emerging from the tests and commentaries are offered throughout the chapter. In particular, the examples of contributors' performance tests scoring criteria (Table 3-3) strongly suggest the influence of teacher theory as discussed in Chapter Two (Figure 2-1). Finally, there is a chapter summary and an annotated list of suggested reading for further understanding.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Definitions of Key Concepts

Decisions made with tests. Tests are used to collect information on learners. This information in tum is used to make decisions (Fulcher, 2010a). Thus, tests are given, and the resulting learner scores interpreted, for a specific purpose. Brown (2005) names four decisions: 1.) "proficiency" or "program-level" decisions, 2.) placement decisions, 3.) achievement decisions, and 4.) diagnostic decisions (pp. 7-8). He defines program-level decisions as learner admission to a program, and other decisions related to comparing learners in one school to learners in other schools. These comparisons may be done for program evaluation purposes (Gorsuch & Griffee, 2018). Placement decisions compare learners within a single program to each other for the purpose of putting learners into appropriate class levels (Brown, 2005, p. 7). None of the tests presented by contributors in this book are used for these program-level purposes.

Achievement decisions have to do with whether learners pass or fail a specific course, and/or what grade they should get. To make such decisions, teachers/testers need to know what learners know and can do in relation to course outcomes. Rather than comparing learners to each other, as with admission and placement, learners are compared to the course outcomes and course content (Gorsuch & Griffee, 2018, p. 43). Finally, diagnostic decisions have to do with whether teachers/testers need to re-teach content, plan more class practice time or tasks, plan additional tutorial sessions, or move on to new content (if learners do well on specific sections of a test, suggesting attainment of particular course outcomes). As with achievement decisions, learners are compared to course outcomes and course content. One difference is that a diagnostic test is given at the beginning of a unit or course, or during a course, whereas an achievement test will be given at the end of a unit or course. See Table 3-1 for a digest of contributors' tests reflecting the language the tests focus on; the decisions made with the tests; and whether the tests are paper and ink CRTs, or performance tests, or both.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 3-1: ***Digest of contributors' tests and decisions made with them***

Contributor(s) and test language focus	Decisions made with test	Test type
Kai Ying Hsu Mandarin Chinese	Learner achievement, and "to investigate how could the authentic materials be integrated into the normal classroom instruction."	Paper and ink CRT and performance test
Dale Griffee English	Learner achievement, to award a course grade, and learner diagnosis.	Paper and ink CRT
Myles Grogan English	Learner achievement, and "to counter a lack of communication experience in English learners in compulsory university EFL classes in Japan."	Performance test
Juliana Jandre and Vander Viana English	Learner achievement, learner diagnosis, and "stimulating students to consider cultural diversity further."	Paper and ink CRT
Ferit Kilickaya English	Learner achievement, and to award a course grade.	Paper and ink CRT
Gisela Mayr English	Learner diagnosis, to better plan an upcoming year's curricuhun, to offer learners feedback, and to introduce the notion of a speaking competency test at the school.	Performance test
Sakae Oneda English	Learner achievement, and "to raise learners' awareness of English collocations."	Paper and ink CRT
Meredith Stephens and Meagan Kaiser English	Learner achievement, and "to foster speaking skills" including fluent use of collocations and leading "learners to process English in its natural order."	Performance test
Beatriz Garcia Glick French	Learner achievement, and to offer students "feedback on their written and pronunciation skills."	Performance test
Annis Shaver German	Learner achievement, and to design a rubric that would not unduly penalize learners for poor "pronunciation, grammar, and affective appearance."	Performance test

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Borbala Gaspar and Margherita Berti Italian	Learner achievement, to offer learners feedback, and foster "opportunities to connect their [students'] personal and /or career interest with the Italian language."	Performance test/Project
Taichi Yamashita Japanese	Learner achievement, to offer learners feedback, and to award a grade	Paper and ink CRT
Irina Drigalenko Russian	Learner achievement, learner diagnosis, and to aid learners' learning through "noticing the conditions that are required for output to be useful in a real-life, organic environment."	Paper and ink CRT and performance test
Yesica Amaya Spanish	Learner achievement "on a lesson based on authentic materials" designed to help learners "work on culture" in addition to language skills.	Paper and ink CRT and performance test
Maria Martinez-Garcia Spanish	Learner achievement, learner diagnosis, and to offer learners feedback not only on individual work, but on group work.	Performance test/Project

Note how contributors cite multiple purposes for writing and using their tests, in addition to creating data on which to base achievement and diagnosis decisions. For instance, Sakae Onoda wants to raise learners' awareness of specified linguistic forms. Myles Grogan intended his performance test to provide opportunities for learner engagement in conversation.

In general, program-level proficiency and placement decisions are made using norm-referenced tests (NRTs), and in some cases, performance tests. Achievement and diagnostic decisions are made using criterion-referenced tests (CRTs) and performance tests. As can be seen in Table 3-1, the tests offered by contributors in this book are, by and large, CRTs and performance tests used for achievement and diagnostic decisions. The decisions to be made using test scores have significant implications for test design. This will be highlighted in the definitions that follow. Test design refers to what skill or knowledge a teacher/tester plans to capture in a test, what test item types are used (receptive/productive, fill in the blank, short answer, etc.), and how a test is administered and scored, among other things.

Norm-referenced tests. There are no norm-referenced test (NRTs) contributions to this book. But the term is defined here because NRTs seem

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

to have a hold on the popular imagination, standing in for what a "real" test should be. This may conceal the fact that NRTs have no real use as classroom tests. Defining NRTs as a concept also helps bring to the fore the issue of what second language tests, whether NRTs, CRTs, or performance tests, are best suited to measure, and what impact that has on practical aspects of test design. Test design affects how a test will look and what learners will experience when they take the test.

Norm-referenced tests (NRTs) are paper and ink or computer administered tests that are designed to capture learners' general second language abilities, or proficiency. In test design terms, NRTs typically use receptive item types where learners read a prompt and then choose an answer, from an array of answers provided. Generally, NRTs use multiple choice items arranged in long subtests arranged around a general theme, such as "listening" or "reading" in the case of *TOCFL* (Test of Chinese as a Foreign Language; Taipei Economic and Cultural Office, 2011), or "listening comprehension," "language form and meaning," and "reading comprehension" in the case of *TOEFL Junior* (Educational Testing Service, 2019b).

As will be seen in Chapter Five, proficiency is a theorized ability to use language for some future, undefined activity (Davies, 1990). The overarching goal of language proficiency test writers is to place large groups of learners with diverse abilities, backgrounds, and circumstances on a single scale. Learners' place on the scale is derived from their scores on the test (to see a scale, look at the x-axis, the horizontal axis, in Figure 3-1). Thus, learners across different institutions, countries, and global regions can take the test and then be compared to each other on a single scale. By design, then, NRTs have no relationship to any given course or program, and thus cannot be used to directly measure learning in a program. Rather, an NRT is related to a norming group, hence the term "norm" in norm-referenced test.

NRTs are developed using large norming groups of learners, again, with diverse levels, backgrounds, and circumstances. Test writers administer large numbers of items in long subtests under general proficiency subtest headings, such as listening proficiency (e.g., Goh & Aryadoust, 2010). The norming group, often numbering in the thousands, takes the test. See Goh & Aryadoust, 2010 for a description of their norming group. Then the test writers mathematically analyze learners' responses to each item, keeping items that function effectively to put members of the norming group on a scale from low to high. In practical terms, test writers want one third of their test items to be hard for the norming group, one third of the items to be of medium difficulty, and one third of the items to be easy. The resulting test

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

made up from the remaining easy-, middle-, and difficult-level items relative to the norming group is intended to create a normal distribution when the norming group's scores are placed into a graph. See Figure 3-1.

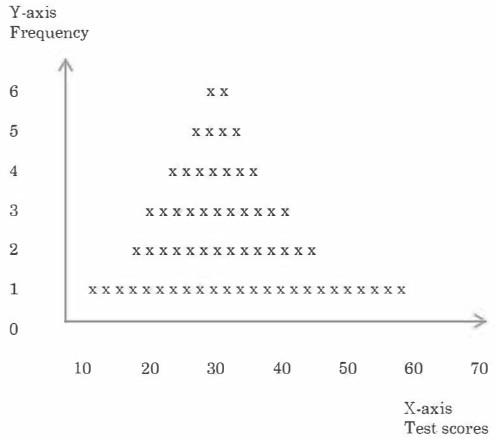


Figure 3-1. **A histogram showing a normal distribution characteristic of an NRT that has a maximum score of 70 (adapted from Gorsuch & Griffee, 2018, p. 44).**

Figure 3-1 shows a normal distribution where a few learners get low scores, many learners get scores in the middle, and a few learners get high scores.

To restate, NRTs are designed to measure proficiency. Proficiency is theorized as a stable trait, or unitary general ability, that is relatively unchanging over time. Many normal distributions are naturally occurring, such as height or shoe size. If one were to collect 100 people together and measure their heights or shoe sizes, the resulting distribution would look very much like Figure 3-1. Again, this points to a view that holds that a general proficiency test, an NRT, has no assumed relationship with specific programs of second language study. Needless to say, to write an NRT is an expensive, time consuming task requiring expertise in psychometrics, and so it is relatively rare to see a large-scale NRT written by working teachers.

What are NRTs good for? Some of the characteristics of NRTs make it useful for program-level decisions, such as admission or placement (Brown, 2005). First, because a well-constructed NRT will result in a normal distribution, test administrators can make a cut point, choosing only the "best" test takers for admission (the upper tail of the distribution in Figure 3-1). Second, that same normal distribution resulting from a placement test

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

given at the beginning of a program can be cut into three parts if one wants three groups roughly ranked by general language ability. Brown (2005) makes an important point, however, that NRTs used for placement ought to be chosen carefully, or designed carefully, so that the test items have some kind of meaningful relationship to a program. Brown uses an example of administering a general grammar proficiency test to learners in a program, and then the futility in attempting to use the resulting scores to place learners in three levels of a speaking class (2005, p. 11). Thus, if learners need to be placed in one of three speaking skills classes, a speaking proficiency test is needed. This, however, would not be a paper and ink or computer administered NRT. A speaking proficiency test would be a performance test, which is discussed below. Gisela Mayr's contribution in this book is a rare example of a locally developed performance test used for placement and diagnosis in her school.

Finally, NRTs are good for comparing learners across programs. Learners taking an NRT in one school can be compared to learners taking the same NRT in another school. In essence learners in both schools are being compared to the norming group that was used to design the NRT. A school administrator or teacher can then use learners' scores to compare their program to other programs, and to refine their advice to graduates as to what schools they may wish to apply to for further study, or what jobs they can seek.

Criterion-referenced tests. Criterion-referenced tests (CRTs) are the common classroom tests and quizzes used by teachers to assess learners' achievement, to provide feedback to learners, to guide review and further classroom practice, and to award course grades (but see additional teacher purposes in Table 3-1). Many of the contributions to this book are recognizably CRTs. They are the everyday paper and ink tests commonly used to test learners on course content. CRTs commonly have four or five subtests with five to ten test items in each subtest. The subtests can be determined by course objective (recognizing and using vocabulary for getting and giving directions, being able to interpret speaker intentions in interactions, inferring meaning from unknown words in reading texts), or by a teacher's conception of language knowledge or skill as it relates to the course objectives (using the correct verb forms in sentences, recognizing a speaker's social status from a listening extract). At one level of practical test design, then, the tying together of a classroom test and the course curriculum finds expression in test item types being grouped together, such as matching, true/false, cloze, or short answer items. With some of the tests in this book, the contributors include one or more subtests on their CRTs that are performance tests, where learners write a paragraph. See Table 3-1 where

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

some contributions are described as "paper and ink CRT and performance test."

It is unlikely that working teachers/testers refer to their quizzes, end-of-unit tests, mid-term exams and final exams as CRTs. Yet their tests have many features in common with a broad, or technical, understanding of CRTs that has emerged thanks to pioneering language testing specialists, working from broad movements in general educational testing.

Criterion and mastery. Two terms are key to defining criterion-referenced tests (CRTs): 1.) criterion and 2.) mastery. Brown & Hudson (2002), two leading theorists and proponents of second language CRTs, define criterion as the specified knowledge or skill a CRT is designed to capture. In terms of classroom tests, then, the criterion is whatever a teacher/tester perceives to be the content, knowledge, or skills to be achieved by learners in a language course.

For instance, in his contribution, Dale Griffee wanted to find out how much vocabulary learners had retained from a textbook chapter, and whether learners could use the vocabulary in sentences. These were specific manifestations of a general course outcome, which was to boost learners' academic vocabulary in English. He also wanted to know: "how they could use their inferencing skills" with long paragraphs in popular news journals. Taichi Yamashita, another contributor, wanted to know if learners could use their listening skills to "identify the location of buildings" on a map, as they had practiced in class. Again, this was a specific manifestation of a general course outcome on developing learners' listening comprehension.

The term mastery has to do with how well learners measure up against the criterion. In other words, how well do they know something they learned on a course? How well can they do something they have practiced in and out of class? Thus, a CRT relates learners to a criterion, hence the term "criterion" in criterion-referenced test. To return to Dale Griffee's vocabulary quiz, it can be inferred that his conception for mastery for vocabulary learning in the course is tied to learners being able to use the new vocabulary in a sentence. There are then two levels to the idea of "mastery." On one level, learners are judged on mastery in terms of how many test items in a subtest they could answer correctly. Griffee notes that if learners got fewer than seven out of ten points on the subtest he would doubt learners had mastery. But on another level, learners are also judged on mastery in terms of what they are asked to do in a CRT subtest. Griffee asked learners to write sentences with the vocabulary they had studied, which is harder to do than learners simply recognizing words and matching them to definitions, or writing definitions. Teacher theory, which is described in Chapter Two, results in deep knowledge of, and sensitivity to,

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

specific learners in specific courses. It is argued here, and elsewhere in the book, that teachers/testers use teacher theory to intuit what the criterion is for a given classroom test, and also what constitutes mastery, probably on multiple levels.

The different decisions teachers make about criterion and mastery, according to their teacher theory, have impacts on practical test design. Teachers/testers use both receptive and productive test item types on their classroom tests, reflecting their grasp of what learners ought to know or be able to do in relation to course outcomes. Within the receptive and productive categorizations, they use a variety of item types, each of which influences what learners experience as they take a test. To further illustrate, see Table 3-2.

Table 3-2: ***Teacher decisions about criterion and mastery expressed in classroom test subtests***

Contribution	Sample subtest, criterion, unit or course objective, and mastery level	Test item type and description
Juliana Jandre and Vander Viana "An English as a Foreign Language Test for Reading, Writing, and Cultural Diversity Awareness for High School Students"	Sample subtest: Subtest 1 (of four), four items Criterion: Learners' reading comprehension of a short text, their ability to analyze the text, and their ability to engage in argumentation and self-reflection. Course objectives: To build learners' reading and writing skills; Stimulation of learners' awareness of cultural diversity. Mastery level: Flexible, although learners should get 60% of items correct overall, either on the test or through course participation.	Test item type: Receptive (reading a text) and Productive, short answer Description: Learners read a short text in English with four words highlighted. Learners answer two text comprehension items in English and then and two meaning analysis/opinion questions in Portuguese.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Irina Drigalenko "A Written and Oral Russian Achievement Test for Beginning College-level Learners"	Sample subtest: Subtest 3 (of seven), three items Criterion: Learners' ability to interpret size and cost Course objectives: Learners engage in output that is useful in a real-life, organic environment. Learners focus on various foods and beverages in Russian, in conjunction to the verbs for "to eat," "to drink," "to buy," and "to sell." They also learn how to count money, compare prices, and shop for food. Mastery level: Learners must get two out of three items on the subtest correct.	Test item type: Productive, short answer Description: Learners read an online table of caviar can sizes and prices and answer three questions in Russian, such as "Which can is the most expensive?"
Yesica Amaya "La Historia de la Pola: An Achievement Test for Original Content-Based Materials for Beginning Learners of Spanish"	Sample subtest: Subtest 1 (of four), ten items Criterion: Learners' word recognition skills Unit objective: Learners develop bottom-up listening skills Mastery level: Learners must get six out of ten items on the subtest correct.	Test item type: Receptive, multiple choice Description: Learners listen to ten sentences read in order from a timeline of la Pola's life events; Learners circle the word they hear from four choices.

Jandre and Viana wanted to capture not only reading comprehension but also evidence of learners' ability to engage in argumentation on culturally sensitive topics brought up in the test text. They designed receptive and productive items for one subtest where learners had to read a text and then

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

answer comprehension questions in English. But then learners also had to argue for positions in response to the text in Portuguese. The contributors adjusted the task requirements (writing in learners' L1s) for the last two items to adjust for an arguably harder task (argumentation). This suggests different levels of planning in their test design where they want to capture learners' performance against their criteria using flexible conceptions of learner mastery.

In Yesica Amaya's test, her interest was to capture learners' word recognition skills. Thus, she designed items that were receptive, which allowed learners to focus on selected words. At the same time, learners had to listen to ten utterances. It was not a short subtest in this regard. Like Griffee, Amaya's conception of mastery had two levels of planning resulting in specific features of test design in relation to her chosen criterion: First, writing fairly easy and yet focused receptive type items for learners to answer; and Second, assigning a relatively large number of items for learners to answer and setting "mastery" at 60% (out of ten items), something less easy.

Test purpose and distributions. Brown and Hudson (2002) further note that a criterion-referenced test (CRT) is "any test that is primarily designed to describe the performance of examinees in terms of the amount they know of a specific domain of knowledge or set of objectives" (p. 5). Thus, when learners take a CRT, their performance is compared to the course content. Learners are not compared to each other. In contrast to the normal distribution an NRT is intended to result in (Figure 3-1), then, CRTs result in non-normal distributions. Figure 3-2 shows the interaction with learners and a diagnostic test, which captures course content learners have not yet engaged with (a pre-test).

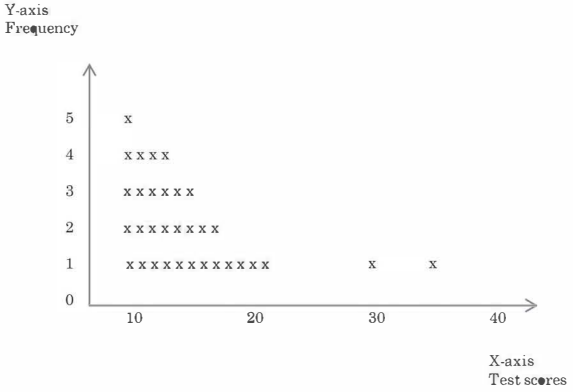


Figure 3-2. **A histogram showing a non-normal distribution characteristic of a diagnostic/pre-test CRT that has a maximum score of 40 (Adapted from Gorsuch & Griffee, 2018, p. 44).**

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Many learners got low scores on the test, while just a few learners got higher total scores. There are two outliers here, one with a score of 30 out of 40, and another with a score of 34. This would not be surprising in a typical language class. The outliers may be heritage speakers of the language, with family members who use the language, or they had study or living experience abroad or additional coursework at another school. Having this information allows teachers/testers to confirm the content they had planned to work on, and alerts them to two learners who may need different or additional treatment than the planned content.

When learners take a CRT at the end of a unit or course (a post-test), and have engaged in the course content, yet another, strikingly different, non-normal distribution would be expected. See Figure 3-3.

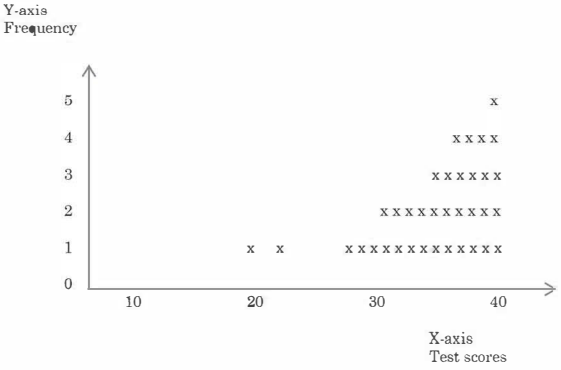


Figure 3-3. **A histogram showing a non-normal distribution characteristic of an achievement CRT that has a maximum score of 40 given as a post-test (Adapted from Gorsuch & Griffee, 2018, p. 45).**

If learners do their jobs by attending class and engaging with the material, and if teachers/testers do their jobs and focus and assist with learners' engagement with the content, and if the materials are adequate, and if the test reflects what learners do in class, many learners should get high total scores. A few outliers/students would be expected in real life. Note one learner who got a 20 and another who got a 22. Perhaps they did not attend class, or had trouble engaging with the content for whatever reason. A teacher/tester may set a cut score of 28 out of 40, which reflects their conception of mastery of the criterion (course content expressed in the test), with the result of two learners failing the test, and the remaining learners

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

passing the test. The purpose of showing the distributions in Figures 3-1, 3-2, and 3-3 is not to suggest teachers/testers must use histograms but rather to show how CRTs (classroom tests) focus on comparing learners to course content. The distribution greatly changes shape depending on when the test is given relative to when learners engage in the course content.

Multiple conceptions of test purposes: What are CRTs good for? On one level classroom tests can be described by the decisions they are used for (outlined above as diagnosis and achievement). On another level, a teacher/tester likely has purpose for a test that is stated in their own terms and needs, such as mid-term exam, final exam, or quiz. Behind these familiar terms are multiple purposes, including evaluation of materials and instruction, reviewing content, motivating students, raising learners' awareness, and offering feedback to learners. For instance, Kai-Ying Hsu reported that her Chinese test was designed "to learn whether authentic materials were effective to promote language learning." She had made a unit using authentic materials and video clips found online, and she wanted to know whether learners could handle the materials and learn new language content. Sakae Onoda intended her English verb collocation test to raise test takers' awareness of words used with high-frequency verbs and to minimize learners' reliance on direct L1 translation of collocations. Many contributors of paper and ink classroom tests reported their tests could be, and were used for learner feedback, but also stated that this was an added-on purpose that was secondary to the primary purposes listed above (see Table 3-1).

Many contributors reported that their classroom tests and quizzes were used to determine course grades, including Irina Drigalenko, Ferit Kılıçkaya, Dale Griffee, and Taichi Yamashita. Some contributors, such as Dale Griffee, spoke specifically how they designed their tests (the number of items and subtests, etc.), and their scoring procedures (awarding weighted scores to test items, etc.) to result in some proportion out of 100 (as in 100%) for ease of interpretation, or due to institutional requirements. Other contributors, such as Irina Drigalenko, simply designed their tests to add up to some conception of 100 points without overt discussion of it in their contribution. See Chapter Seven.

Performance tests. Performance tests are tests where learners do a task using the second language, and then some aspects of their resulting language use are scored by a teacher/tester (Gorsuch & Griffee, 2018; Norris, Brown, Hudson, & Yoshioka, 1998). Like CRTs, teachers/testers use performance tests to get a handle on learners' achievement, to give feedback to learners, and to award grades. And, as was seen in Table 3-1, teachers use performance tests for other reasons, including motivating learners, helping learners become more self-aware, and encouraging learners' decision making and

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

connections to new content. Quite a few contributions to this book are performance tests. Some are simply a subtest of a longer paper and ink classroom test (see for instance, Hsu, Drigalenko, and Amaya). Other contributions are stand-alone performance tests in which learners' second language use and production comprise the main activity of final projects, and mid-term and final exams (see for instance Grogan, Stephens & Kaiser, Gaspar & Berti, and Martinez-Garcia).

Performance tests are prized by teachers because they may be authentic and more closely related to real-world uses of language. This can be seen as motivational to learners. Further, many second language programs are adopting aims and objectives specifying learners' development of speaking, writing, and interacting for purposes of communication. Two contributors cite these objectives as primary drivers of their making and using the performance tests they offer in this book (Annis Shaver teaching German in the U.S. and Maria Martinez-Garcia teaching Spanish in Korea). Both the ACTFL *Guidelines* (American Council on the Teaching of Foreign Languages, 2012a) and the *Common European Framework of Reference* (*CEFR*, Council of Europe, 2001, 2018) promote communicative language use activities and tasks for second language learners in classrooms (see for instance Gisela Mayr, who cites *CEFR* as a significant driver of her decision to make and use a speaking performance test). Making and using performance tests has become a professional necessity.

Like CRTs, it is unlikely that teachers/testers refer to their performance tests as such. If the sample presented in this book is any indication, performance tests vary greatly in form and process. The simple or multi-layered, multi-occasion tasks learners are asked to do, and the scoring systems then used by teachers, are diverse. The teacher theory model (Figure 2-1) predicts this, with its components of teacher background, past education and workshops, current teaching, and institutional context. Yet, to make sense of this diversity, it is useful to consider performance tests as they are formally described in the testing literature.

Test tasks. There are two terms key to defining performance tests in the testing literature: Test task and scoring criteria. A performance test task is an action or activity learners engage in to produce a score on a performance test. Task is a term coming from communicative language teaching in which learners engage in some activity that has an end goal, while using the second language in an extemporaneous fashion (Gorsuch & Griffee, 2018, p. 342). The term "extemporaneous" means that learners are not reciting a memorized dialog, or writing a memorized or contrived sentence or two. Rather, learners make an assessment of what a test task requires them to do, and use and alter what they plan, monitor, say, or write, accordingly. A

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

performance test task could look like learners of Spanish writing a postcard to their friends after a classroom unit about an early 19th century Colombian heroine (see Yesica Amaya's contribution), or learners of German engaging in group discussions on authentic short stories they have read (see Amis Shaver's contribution).

There remains robust and sustained discussion in second language testing as to what constitutes good performance test tasks. This is an important discussion in that whatever we ask learners to do on a test affects the score a learner gets. Teachers/testers then use the scores to make significant decisions about their own future teaching and testing, and about learners' course grades. One discussion centers around the authenticity of a performance test task, with the idea that such tasks should reflect as much as possible real-world tasks learners might be expected to perform in future (Patton, 2011). Jones (1985) offers a still-useful and still-used taxonomy of performance test tasks that elaborates this concern with capturing real-world tasks, where he describes planning tasks based on actual observations done in a workplace. Other scholars point out that good performance test tasks can emerge from pedagogical work (Bachman, 2002; Hawkes, 2012). Many teachers/testers ask learners to write essays or engage in group projects, for instance. Such learner activity occurs in classrooms, but not necessarily in workplaces.

There is one consensus of what constitutes an adequate test task—the task must be consciously and credibly related to the curriculum and objectives of a course (Gorsuch & Griffee, 2018; Patton, 2011; Purpura, 2016). This contrasts with general proficiency tests, that, while appearing to be a performance test, are not connected to any particular language curriculum. An example is the ACTFL OPI spoken proficiency interview (American Council on the Teacher of Foreign Languages, 2012a; Center for Open Educational Resources on Language Learning, 2010)(see Chapter Five on Communicative Competence and Language Use Description Frameworks). Such general proficiency tests must be seen as most suitable for use for program-level decisions, much like NRTs, but not for classroom decisions, such as achievement or diagnosis (see Luoma, 2004). See Table 3-3 below for descriptions of performance test tasks offered by contributors to this book.

Scoring criteria. Scoring criteria are descriptions, set to numeric scales (1, 2, 3, etc.) that learners' performances are compared to. In this sense, performance tests are like criterion-referenced tests. Learners are compared to the criteria, and not to each other. Scoring criteria (often called "rubrics" by contributors and teachers/testers at large) come in two basic forms. They are either analytic or holistic (Brown, 2005; Gorsuch & Griffee, 2018;

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Luoma, 2004). Analytic scales are a collection of two or more scales, with each scale representing some aspect of a learners' performance the teacher/tester wishes to focus on. A holistic scale has a scale set against a single description that is longer, more complex, and more general. For instance, Yesica Amaya uses an analytic five point scale with three criteria: "Organization," "Content," and "Grammar." In her scoring criteria, the descriptions appear under each point on the scale, and are short. A score of 2 on Organization on a learner's postcard writing reads "The organization of the events does not follow a logical sequence or are not enough." A learner getting a 3 on Content would get: "The text explains some of the events about Pola's life with their own words and there are only a few inaccurate information." A learner getting a 4 on Grammar would get: "Effective use of grammar but some mistakes." Kai-Ying Hsu and Myles Grogan appear to be the only contributors who use a holistic scoring scale. On Kai-Ying Hsu's test, learners' do a paragraph writing task. There are four points on the scale from 0 – 3. For instance, a learner getting a score of 2 gets: "Uses some appropriate vocabulary and grammar to describe both characters and events."

Scoring criteria of contributors to this book. Here we move beyond how the traditional testing literature defines and describes performance tests. The following section points out what working teachers/testers themselves say about the scoring criteria they find, adapt, make, and use. Needless to say, there is great variation in teachers'/testers' scoring criteria within this book, and one would think, out in the working world. To attempt to distill this variation almost seems to unnecessarily reduce these nuanced, externalized evidences of teacher theory (Figure 2-1). Yet some sense should be made of them, however clumsily. The discussion here focuses on two ways this variation might be understood: 1.) Whether a teacher/tester makes or uses an analytic scale or a holistic scale; and 2.) The sources of theory teachers use to find, adapt, make, and use scoring criteria.

Why analytic and why holistic: Practical exigencies. Teachers'/testers' decisions to use analytic and holistic scores, defined earlier, may be shaped by whether their performance test is a stand-alone test with some import, or just one subtest among others in a criterion-referenced test (CRT) which is otherwise in a paper and ink format. Stand-alone performance tests used for classroom decisions may result in longer speech, interaction, or writing samples, which, being complex performances, would lend themselves to analytic scoring. Performance tests for speaking or writing that are just one subtest out of several in an end-of-semester criterion-referenced test may lend themselves to holistic scoring. There may not be as much of a speaking or writing sample to score.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

A second shaping force over analytic versus holistic decisions may be shaped by teacher/tester time constraints, both during and after the performance test. If teachers/testers are rating speaking performances in real-time conditions, they are truly pressed for time and attention. It may be tempting to use a holistic scale with four or five points with a single, albeit long, description under each point on the scale. Even if a performance test results in writing samples, which can be scored at leisure after the fact, teachers/testers may still be constrained by time in terms of how long they need to invest in scoring the writing samples, particularly if they are working alone with large numbers of students.

Many contributors who made performance tests commented on time constraints. Yet, interestingly, at least seven persisted in using analytical scales, citing a number of strategies for completing complex scoring using analytic scales in limited time/attention conditions. This suggests the contributors saw analytic scoring scales as having real value. Gisela Mayr noted that two examiners participated in her performance test. One teacher/tester moderated the paired-student performance test and the other scored while the test was underway. Amis Shaver, who had to work alone, said she wrote her general impressions on her scoring criteria sheets in real time, and then later after the test made notes and did the actual scoring. Irina Drigalenko persisted, and then coped by using scoring criteria for essay writing that were in "common use," such as "fluency" and "vocabulary range." Gaspar and Berti noted that making their scoring criteria *more* detailed actually helped them deal with time pressure during real time scoring conditions.

Scoring criteria as teacher theory. This persistent use of analytic scales points to the second focus of this section: The sources of theory teachers use to find, adapt, make, and use scoring criteria. As will be described in Chapter Five, one way to look at theories is through the High Middle Low (HML) model. For the sake of this chapter, it is postulated that many scoring criteria scales in the contributions to this book, are, by and main, the result of teacher theory, or "low" theory (The "L" in HML). Teacher theory is localized, private, efficient, purposeful, and bound to a specific classroom context (see Figure 2-1). As will be seen in Table 3-3 below, some contributors' theories were informed by middle-level theories (Amis Shaver's use of Grice's maxims) and high-level theories (Communicative Competence). Yet the middle and high theories, when they were rarely mentioned, seemed to play out in specific classroom contexts and thus were mediated by low-level teacher theory (Figure 2-1). The issue here may be to identify what aspects of teacher theory most clearly shaped contributors'

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

choices of scoring criteria. See Table 3-3 for descriptions of test tasks and scoring criteria, and their stated sources, in this book.

Table 3-3: ***Test tasks, scoring criteria, and stated sources of performance test contributions***

Contributor(s) and name of test	Test task and source	Scoring criteria and source
Kai-Ying Hsu "A Chinese Achievement Test for Intermediate College-level Learners"	Task: A paragraph writing task. Learners are asked to describe characters and events in a video clip they have been watching in a classroom unit. Details: The task comprises the last of four subtests on a CRT paper and ink test. Source: From a teaching coordinator who "often designs tasks for students with…videos."	Scoring criteria: Holistic with four points on a scale. Contributor notes: "Since students were only asked to compose a paragraph, there were not many constructs to capture." Example: 3 = Successfully uses appropriate vocabulary and grammar to describe both characters and events with details. Source: Communicative competence: "To probe learners' discourse (textual) competence, I designed a writing task…I wanted learners to show they could synthesize grammatical knowledge they know (e.g. vocabulary, grammar) and express a textual message…based on what they observed from the film clip"
Myles Grogan "A Simple Speaking Test for an English-Language University Communication Class"	Task: A group discussion with four or five learners on selected, familiar topics previously practiced. The topic for the test is decided by coin flip. Details: The test is one of two such tests in a semester.	Scoring criteria: Holistic with five point range "categories" of 0-2 points, 3-4 points, 5-6 points, 7-8 points, and 9-10 points. There is a one to two sentence description for each category, along with five to seven bullet points describing specific conversational behavior ("conversation features"). There is also an objective score given to teach student with 1 point for each active verb a learner says.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

	Source: Tasks and topics in "compulsory texts" including typical "classroom skills" sections found in many textbooks. The group discussion format is also a regular classroom activity.	Example: For the 5-6 point category, the prose description reads: "A score of 6 indicates a basic level of participation, with a small range of skills used, but some limited success in communication. Typically, a conversation at this level sees every member scoring, but experiences occasional breakdowns or remains limited in depth." Two of the five conversation feature bullet points state: "A student may move off topic" and "A student often focuses on particular speakers and excludes others." Source: "Conversation with a colleague who had seen a conference presentation in which each sentence was given a point" and "A needs analysis based on classwork."
Gisela Mayr "A Speaking Skills Test for High School Learners of English in Southern Tyrol"	Task: A monologic talk. Learners are given a short text written at their level, and then they summarize the text orally. Then they orally answer comprehension and opinion questions about the text. Details: The task is the first of two phases in the performance test: A monologic talk, and then a dialogic role play task done with a classmate.	Scoring criteria: Analytic with five "competence" areas, each with one to three "descriptors" (criteria) with five points on a scale: "Is not true," "Is rarely true," "Is sometimes true," "Is often true," and "Is always true." Example: One competence area is "Cleanliness of contents and thematic representation." Under this are two criteria, "The learner can represent their own point of view clearly and cohesively" and "The learner can expose the topic in a complex and differentiated way." Source: Common European Framework of Reference (*CEFR*) (Council of Europe, 2001). She

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

	Source: Her thoughts on what "students have to do in real life situation…a formal situation, e.g., study, university, or work"	also cites the school curriculum, which is informed by the South Tyrolean Framework for Foreign Language Teaching. She notes this document, and the school curriculum, are derived from *CEFR*.
Meredith Stephens and Meagan Kaiser "Providing an Oral Summary of a Written Text as a Mid-semester and Final Test"	Task: Learners give a monologic five-minute oral summary of a text they have read in the preceding weeks, with ten minutes preparation. Details: The test is a mid-term and final exam, modeled after weekly vocabulary and pair summary tasks learners do throughout the semester. Sources: Adapted from a writing textbook that had a section on writing a summary. Also ten years of coursework aimed at encouraging learners' use of common collocations in speaking.	Scoring criteria: Analytic with five criteria with 1 4 points possible for each ("Introduction" 1 point, "Collocations" 4 points, "Pace" 2 points, "Coherence" 2 points, "Conclusion" 1 point). Each criterion has a one to four sentence description. Example: Collocations: "Has the student remembered the collocations correctly? Are the collocations used correctly in the story? Does the place where a collocation is used make sense in the story? Has the student correctly understood the meaning?" Pace: "Is the student able to tell the story fluently without an overabundance or hesitation or rush?" Source: Ten years of experience teaching speaking courses and noticing that a major area of "difficulty" was using collocations.
Beatriz Garcia Glick "An Oral VoiceThread Test for First-semester French Language Learners in the U.S. University"	Task: A monologic talk on a theme recorded on an online recording and presentation application, with learners working from home. Learners must produce two "paragraphs" of talk with an accompanying slide	Scoring criteria: Analytic in form. There are six untitled points on a scale (0, 1, 2, 3, 4, 5), and two untitled criteria. Each point in a given criteria contains a one or two sentence descriptor. Example: Learners earning a "3" on the first (untitled) criteria gets: "Slide with 5-6 sentences in French." On the second criteria: "Recording with 5-6

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

	of written work using guided prompts and specified sentence structures. Details: The task is one of two VoiceThread assignments in a semester. Source: "Tasks that are relevant and practical in every day life. I also follow the ACTFL guidelines."	comprehensible sentences in French." Learners getting a "5" would get: "Slide with more than 7 sentences in French. No misspelled words" and "A full one minute of comprehensible spoken French." Source: Textbooks the contributor has used. Author also cites the ACTFL *Guidelines* "and descriptors" (American Council for the Teaching of Foreign Languages, 2012)
Annis Shaver "A Speaking Fluency Test for Intermediate-level German Using a Rubric Based on Grice's Conversational Maxims"	Task: Learners engage in small group discussion based on ten prompts, in German, from a German short story they have read. Learners may have a vocabulary notecard. Details: Ten such discussions take place during two semesters. Source: Discussion prompts come from the textbook and are modified "to match students' communicative competence." The discussion format comes from contributor's thinking on "real-world language use which require them to express opinions and consider the opinions of others."	Scoring criteria: Analytic with four criteria, and a five point scale for each criterion. Each criterion has a two to three sentence explanation. Examples: The four criteria are: "Student contributes enough information to the discussion" and "Student's contribution is relevant to the discussion," and "Student contributes clearly: pronunciation, grammar" and "Student's contribution clearly answers or addresses the question." A sample explanation of the 'student's contribution' criteria is: "The Maxim of Relevance is assessed through the relevance of the student's responses to the topic of discussion. Responses should not be 'off topic' without clear or sufficient connection to the discussion." Source: Grice's Conversational Maxims (Grice, 1975)

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Borbala Gaspar and Margherita Berti "A Multiliteracies-oriented Project-based Assessment for Intermediate Foreign Language Italian Classes"	Task: A project where learners electronically upload successively more elaborated installments and finish with a live oral presentation. Details: The project has five steps: 1.) Research proposal draft, 2.) Revised research proposal, 3.) Draft of presentation slides, 4.) Revised presentation slides, and 5.) Ten minute oral presentation. The process takes place throughout the semester. Source: A multiliteracies model in which suggests learners do best in when engaged in successive learning experiences with "situated practice," and work with many textual forms including "graphics, music, sound, print, images." Contributors were also inspired by what they thought learners would do in real life with Italian.	Scoring criteria: Analytic, with a different set of criteria for each of the five steps of the project. Each set of criteria includes 7 10 criteria with three levels each (5-4 points, 3-2 points, 1-0 points). Each level for each criteria has a 2 4 sentence descriptor. Examples: The "Presentation slides draft" criteria include "Use of primary sources," "Content depth and transitions," and "Content accuracy and comprehensibility." A descriptor for a 3-2 point performance on "Use of primary sources" reads: "In the slides it appears that the student partially synthesized research. Although critical thinking is not evident, the student somehow connected resources together. The use of quotes is minimal and all resources are listed on the references slide." Source: The contributors wanted to be sure learners included required elements of each stage of the project they uploaded. In line with a multiliteracies model of learning, they intended learners to become experts on their project topics.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Irina Drigalenko "A Written and Oral Russian Achievement Test for Beginning College-level Learners"	Task: There are two tasks, one oral and one written. For the oral task, learners write and perform a dialog based on a template. For the written task, learners can either write a short essay about their food preferences to a Russian homestay family, or they can write an original dialog with a Russian roommate about a meal the learner plans to make with his or her roommate. Details: The oral and written tasks appear as the first and last of four subtests on an end-of-unit achievement test. Source: The contributor's idea that learners should be creative with language to prepare for a real-life situation (homestays and study abroad). Contributor also observed learners participating in role plays in class and got the idea to offer an option to write a dialog.	Scoring criteria: The oral and written task are virtually the same except for a few minor differences in criterion title and descriptor wording. Both are analytic, with a supplemental holistic scale with descriptors that incorporate concepts from each of the five analytic criteria. Criteria are weighted, and each has five scale points: 6 = "improvement needed," 7 = "Satisfactory," 8 = "Good," 9 = "Very good," and 10 = "Excellent." Examples: For the oral task, the criteria include; Fluency/vocabulary usage (25% of score), Language control/structure (20% of score), etc. A learner getting an "8" on Fluency/vocabulary usage would be: "Good" and "Occasionally lacks basic vocabulary, but generally good usage (appropriate for this level)." For the written task, the criteria include: Content quality (25% of score), Language control and structure (20% of grade), etc. A learner getting a "9" on Content quality would be: "Very good" and "Good treatment of skit/topic." Source: The contributor used criteria that "are more or less common for language essay grading." She also wanted to create points on a scale that emphasize the positive and are more useful to learners as feedback.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Yesica Amaya "La Historia de la Pola: An Achievement Test for Original Content-based Materials for Beginning Learners of Spanish"	Task: Learners write a short letter to a friend about what they learned about Policarpa (the subject of the materials). Details: The task is the fourth subtest of a low-stakes ungraded achievement test. Source: What the contributor was teaching in another class, which included an e-mail task.	Scoring criteria: Analytic with three criteria and give levels each. Each level is simply a score of 1, 2, 3, 4, or 5. Descriptors for each criterion appear under each score. Examples: The three criteria are "Organization," "Content," and "Grammar." A learner getting a "4" on Organization gets: "Generally organized with only a few mistakes in the organization of events." A learner getting a score of "4" on "Content" would get: "The text includes most of the events about Pola's life with their own words only a few inaccurate information." Source: Material from a testing course she was taking. She wanted criteria to reflect what she was interested in in terms of learners' learning achievement of the materials she created.
Maria Teresa Martinez-Garcia "A Final Project Performance Test for a Spanish Conversation Class at a Korean University"	Task: Learners prepare a Spanish-narrated video in small groups; They then present the video to the class. The topic was: What does your university mean to you? Source: Attendance at an online institute on "Fundamentals of Project-based Language Learning" offered by the National Foreign Language Resource Center. The institute stressed finding topics	Scoring criteria: Analytic with five criteria with four point scales for each. Each point on the scale is oriented to a letter grade, as in "Above expectations/A" and "Meet the expectations/B," etc. Examples: Criteria are "Vocabulary" and "Grammar" and "Fluidity" and "Comprehension and production" and "Originality." Each criterion has a one to two sentence descriptor under each point on the scale. A learner getting a "B" on "vocabulary" gets: "Employs appropriately and quite effectively the required vocabulary." A learner getting a

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

	of personal interest to learners. At the time the contributor made this test, the students' university was under renovation, and the library, elevators, etc., were not available for a significant amount of time.	"B" on comprehension and production gets "Good understanding of the purpose and task, and appropriate use of the vocabulary/grammar. There is communication between the members of the group and the audience." Source: The online institute she attended, previous experience as an instructor in the U.S., online sources, requirements of the speaking section of the DELE B1 (Institute Cervantes, 2019), and *CEFR* B1-level descriptors (Council of Europe, 2001)

Chapter Summary

In this chapter, four key concepts were defined: Decisions made with tests, norm-referenced tests (NRTs), criterion-referenced tests (CRTs), and performance tests. As CRTs and performance tests are what contributors offered this book, this chapter served as an introduction to these contributions. The definitions and the contributions taken together form a foundation to understand the important and largely unseen educational roles that CRTs and performance tests perform at the classroom level. CRTs and performance tests were portrayed as commonly used to make achievement and diagnostic decisions, with implications for both learners and teachers. With such decisions, learners are compared to content, and not to each other, and the terms of criterion and mastery were introduced and defined to underscore this important difference.

A closer inspection of CRT contributions to this book also revealed additional teacher purposes for CRTs, unrelated to awarding grades, such as motivating students or raising learners' awareness on particular language points. For performance tests, the terms test task and scoring criteria were highlighted, and the terms were used to explore the real diversity in the performance test contributions to this book. The Teacher Theory Model (Figure 2-1) predicts this diversity when one considers the contributions of teacher background, past education and workshops, current teaching, and institutional context to teachers' day to day, principled decisions about tests.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Further Reading

The following are about criterion-referenced tests (CRTs) and performance tests as they are described and discussed in the testing field, and the second language teaching field.

On CRTs

Brown, J.D. & Hudson, T. (2002). *Criterion-referenced language testing.* Cambridge: Cambridge University Press.

Frain, J. (2009). A comparative study of Korean university students before and after a criterion-referenced test. Unpublished M.A. Thesis, University of Southern Queensland, Australia.

General Testing Books with Sections on CRTs

Fulcher, G. (2010). *Practical language testing.* London: Hodder Education.

Gorsuch, G. & Griffee, D.T. (2018). *Second language testing for student evaluation and classroom research.* Charlotte, NC: Information Age Publishing.

On Performance Tests

McNamara, T. (1996). *Measuring second language performance.* London: Longman.

Norris, J.M., Brown, J.D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments.* Honolulu, HI: University of Hawaii Press.

General Testing Books with Sections on Performance Tests

Gorsuch, G. & Griffee, D.T. (2018). *Second language testing for student evaluation and classroom research.* Charlotte, NC: Information Age Publishing.

On Test Tasks for Performance Tests

Douglas, D. (2000). *Assessing languages for specific purposes.* Cambridge: Cambridge University Press.

Hawkes, M. (2012). Using task repetition to direct learner attention and focus on form. *English Language Teaching Journal, 66*(3), 327-336.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Jones, R.L. (1985). Second language performance testing: An overview. In P.C. Hauptman, R. LeBlanc, & M.B. Wesche (Eds.), *Second language performance testing* (pp. 15-24). Ottawa: University of Ottawa Press.

On Scoring Criteria for Performance Tests

De Silva, R. (2014). Rubrics for assessment: Their effects on ESL students' authentic task performance. *Center for English language communication 4th symposium proceedings.* Singapore: National University of Singapore. Available: http://www.nus.edu.sg/celc/research/books/4th%20Symposium%20proceedings/19%29.%20Radhikda%20De%20Silva.pdf

Fulcher, G. (2010). *Practical language testing.* London: Hodder Education. (pp. 208-215)

Jeong, H. (2015). Rubrics in the classroom: Do teachers really use them? *Language Testing in Asia, 5*(6). DOI 10.1186/s40468-015-0013-5

Kuiken, F. & Vedder, I. (2017). Functional adequacy in L2 writing: Towards a few rating scale. *Language Testing, 34*(3), 321-336.

Upshur, J. & Turner, C. (1995). Constructing rating scales for second language tests. *ELT Journal, 49*(1), 3-12.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Chapter Four

The Tests

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

A Chinese Achievement Test for Intermediate College-Level Learners

Kai-ying Hsu
Texas Christian University

Introduction

At the time of this writing, I was an M.A. candidate in an applied linguistics graduate program at a university in the south central United States. I was supported as a Chinese (Mandarin) instructor of record. My research interests continue to be second language testing, Computer-mediated communication (CMC), and second language acquisition. This test was written as a follow-up project for Chinese-language authentic materials that I designed for a graduate course I was taking. During that semester, I made the materials and piloted the test. The materials, the pilot test, and eventually, the test presented here, were based on a film clip I found online called *Opening cooking scene - "Eat Drink Men Women" (1994).* During the materials development and pilot stage, I administered the test with ten Chinese language learners from my school, who volunteered to participate. Participants took the pilot test in our language learning lab. For the pilot, the participants were divided into two contrasting groups: masters and non-masters. The master group consisted of three intermediate and advanced Chinese language learners, whereas the non-master group was made up of seven beginning Chinese language learners. The idea was that the master group would do well on the test, and the non-master group would not do well, if the test were written at the right level.

For the second stage, I revised the test items that did not function well in the pilot test. The revision included reducing the number of the test items, modifying the test instructions, and editing out distracting background noise of listening abstract for the Cloze subtest. The revised test was peer-reviewed by my testing course instructor. Next, I administered the revised test as a pre- and post-test with seven low intermediate Chinese language learners. The second stage took five class

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

hours over two weeks, including three and half hours of instruction and one and half hours of data collection (pre-test, post-test, and questionnaire). During the instruction, students engaged in classroom activities based on the learning materials (an authentic film clip, lesson slides, and a handout), which aimed to prepare students to succeed in the test. The content was congruent with the test content. All learners received bonus points on their final grades as a reward to engage in the pilot test.

The Test

A Chinese Achievement Test for Intermediate College-level Learners

中文名字: _____ Grade: / 20

Authentic materials: The film clip: (1.) *Opening cooking scene - "Eat Drink Men Women" (1994)* retrieved from youtube video: https://www.youtube.com/watch?v=1-2QBYKI8LU&list=PLUmzmsPapCYtnS33QC5GP2DLcktk_fun_

A. Fill in the blanks (5%)

The following are seven kinds of Chinese cooking methods.
How did the male actor prepare the following food? Please fill in the blanks based on what you see in the film clip. You will only use each choice once. You will watch the clip twice.

炸 zhá 炖 dùn 包 bāo 蒸 zhēng 炒 chǎo 酿 niàng

1. 他_____了鱼。
2. 他_____了一盘菜
3. 他_____了一碗猪肉
4. 他_____了饺子
5. 他_____了一只鸡

Note:
1. 鱼[yú]:fish
2. 猪肉[zhūròu]:pork
3. 鸡[jī]:chicken

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

B. Cloze (5%)

Listen to the script and fill in the missing words with Chinese characters or pinyin. You will hear it three times.

亲爱的上帝，求你_____我的爸爸和两个_____，使我们全家人都能认识你，现在我要感谢你，赐我们这丰盛的_____，再一次的_____你，让我们一家和乐团聚在_____，祷告感谢奉主耶稣的名，阿门！

Note:

1. 赐[si]:bestow 2. 丰盛[fēngshèng]:abundant 3. 奉[fèng]:in the name of...

C. Short questions (7%)

Answer the following questions in Chinese language (pinyin or Chinese characters) based on what you observe from the film clip. You will watch the clip once. Please note 2 roles are required.

	●1:他是谁？(please write down his role in the clip.)	●2:他在做什么？
1.		

	●2:她是谁？(please write down her role in the clip.)	●2:她在做什么？
2.		

3. 他在哪儿？

_____

4. What are some cultural aspects familiar and unexpected to you? That is to say, what are the cultural similarities and differences between Chinese culture and American culture? You can answer in English or Chinese. If possible, add as much Chinese character or pinyin as you can based on what you learned from the lessons.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Cultural similarities	Cultural differences

D. Task (3%)

You are the script writer. Based on what you know and observe from the film clip, write a short paragraph in either Chinese character or pinyin to describe the major characters and events in the film clip as if you're going to tell the story to a friend or your audience.

_____
_____
_____
_____

Table 4-1: ***Scoring scale and criteria for Task D***

Scale	Criteria
0	Poorly uses appropriate vocabulary and grammar to describe both characters and events.
1	Rarely uses appropriate vocabulary and grammar to describe both characters and events.
2	Uses some appropriate vocabulary and grammar to describe both characters and events.
3	Successfully uses appropriate vocabulary and grammar to describe both characters and events with details.

English Translation and Answer Key of the Test

中文名字: _____ Grade: /20

Authentic materials: The film clip: (1.) *Opening cooking scene - "Eat Drink Men Women" (1994)* retrieved from youtube video: https://www.youtube.com/watch?v=1-2QBYKI8LU&list=PLUmzmsPapCYtnS33QC5GP2DLcktk_fun_

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

E. Fill in the blanks (5%)

The following are seven kinds of Chinese cooking methods. How did the male actor prepare the following food? Please fill in the blanks based on what you see in the film clip. You will only use each choice once. You will watch the clip twice.

炸 zhá *(fry)* 炖 dùn *(stew)* 包 bāo *(wrap)* 蒸 zhēng *(steam)* 炒 chǎo *(stir-fry)* 酿 niàng *(brew)*

6. 他__炸__了鱼。 (*He __fried__ fish.*)
7. 他__炒__了一盘菜 (*He __stirred-fried__ a plate of vegetable.*)
8. 他__蒸__了一碗猪肉 (*He __steamed__ a bowl of pork.*)
9. 他__包__了饺子 (*He __wrapped__ dumplings.*)
10. 他__炖__了一只鸡 (*He __stewed__ a chicken.*)

Note:
1. 鱼[yú]:fish 2. 猪肉[zhūròu]:pork 3. 鸡[jī]:chicken

F. Cloze (5%)

Listen to the script and fill in the missing words with Chinese characters or pinyin. You will hear it three times.

亲爱的上帝，求你__祝福__我的爸爸和两个__妹妹__，使我们全家人都能认识你，现在我要感谢你，赐我们这丰盛的__晚餐__，再一次的__谢谢__你，让我们一家和乐团聚在__一起__，祷告感谢奉主耶稣的名，阿门！

Dear God, please __bless__ my father and my two __sisters__, and make my whole family to know you. Now I would like to give my thanks to you for this terrific __dinner__. And __thank__ you once again for bringing our family __together__ with happiness. In the name of Jesus Christ. Thank you. Amen!

Note:
1. 赐[sì]: bestow 2. 丰盛[fēngshèng]: abundant3. 奉[fèng]: in the name of...

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

G. Short questions (7%)

Answer the following questions in Chinese language (pinyin or Chinese characters) based on what you observe from the film clip. You will watch the clip once. Please note 2 roles are required.

	●1:他是谁？ *Who is he?* (please write down his role in the clip.)	●2:他在做什么？ *What is he doing?*
1.	他是爸爸。*He is a father.* / 他是厨师。*He is a cook.*	他在做晚餐。*He is cooking dinner.* /他在做饭。*He is cooking.* /他给他的女儿做菜。*He is cooking for his daughter.*

	●2:她是谁？ *Who is she?* (please write down her role in the clip.)	●2:她在做什么？ *What is she doing?*
2.	她是姐姐。*She is an elder sister.* / 她是女儿。*She is a daughter.*	她在祷告。*She is praying.* / 她在吃饭前谢谢上帝。*She is thanking God before the meals.* /她在吃中国菜。*She is eating Chinese dishes.*

3. 他在哪儿？ *Where was he?* 他在厨房做晚餐。*He is cooking in the kitchen.* / 他在家做饭。*He is cooking at home.*

4. What are some cultural aspects familiar and unexpected to you? That is to say, what are the cultural similarities and differences between Chinese culture and American culture? You can answer in English or Chinese. If possible, add as much Chinese character or pinyin as you can based on what you learned from the lessons.

Cultural similarities	Cultural differences
• 吃类似的食物。Eat similar food, such as fish, chicken and pork. • 家族聚餐。Family dinner. • 在饭前祷告。Praying before meals.	• 下厨的方式不同。The cooking methods for preparing the meals is different. • 餐桌的摆设不同。The table setting is different.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

• 家人等所有人都到了才开吃。 Family waits for everyone before eating.	中國人從頭開始準備食材，而美國人購買已經備好的食材做飯。 Chinese people prepare ingredients from scratch whereas American people buy already prepared food to cook.

H. Task (3%)

You are the script writer. Based on what you know and observe from the film clip, write a short paragraph in either Chinese character or pinyin to describe the major characters and events in the film clip as if you're going to tell the story to a friend or your audience.

爸爸一早忙着在厨房做饭给她的女儿。他做了很多菜，比如：炸鱼，炖鸡汤，和饺子。煮了一整天，他终于煮好了，他坐下来和三个女儿一起吃晚餐。在她女儿做了饭前祷告以后，他们一起享用美好的晚餐。

Dad was busy cooking in the kitchen for her daughters in the early morning. He made a lot of dishes, such as fried fish, stewed chicken soup, and dumplings. After cooking all day long, he finally finished cooking. He sat down and had dinner with his three daughters. After her daughter prayed before the meal, they enjoyed a wonderful dinner together.

Contributor's Questionnaire Responses

Section One: Test Planning and Writing

Why did you write the test? What were the purposes of the test?

The purpose of the overall project (materials and the test) was to investigate how could the authentic materials be integrated into the normal classroom instruction. The purpose of the test was to learn whether the authentic materials were effective to promote language learning. Moreover, the project aimed to find out students' perceptions on learning Chinese language with authentic materials.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

How did you decide how many subtests to write? Was there some correspondence between a subtest and a course objective, or was there some other reason to create the subtests you did?
The main consideration was the time constraint. I planned that students will be able to finish the test in around 25 minutes to half an hour. And I found the subsets including matching, cloze, short questions, and a task were helpful to check learners' understanding on the basic concepts and demonstrate their abilities to apply the concepts in a task. Also, it was reasonable for students to complete it in the allotted time.

Did you write as many items or test tasks as you needed, or did you write more than you needed, then pare the number down? How did you decide which items or test tasks to keep? How did you decide which items or test tasks to discard?
In the stage of designing the test, I tried to write as many items as appropriate and related to the course objectives. After finishing the draft of the test, I showed it to another knowledgeable person, to add or eliminate the items that did not function well.

How did you decide how many items to write for each subtest?
I referred to a testing book I was studying. For instance, it stated that it is better not to write more than eight to ten items in a matching type subtest since it may cause overly heavy cognitive load and confusion for learners.

How did you decide how many test items to write in total?
For most subtests, I wrote about six items, which I found appropriate and reasonable. Since there are four subtests, including one performance task, the final test consisted of 18 items in total.

Did you have one version of your test, or did you create a second equivalent version?
I had one version of my test.

Were you concerned at how long the test would take to administer?
Yes, since I borrowed time from students' learning schedules to carry out the project. I tried not to occupy too much time as it might affect learners' normal learning schedule as indicated on the syllabus.

Were you concerned at how long the test would take to score?
Not really. It did not take long to grade.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Were you concerned how you might use the test items themselves for learner feedback?
I did not intend to use the test to offer students feedback on their learning. Since the purpose of the test was to learn whether the authentic materials were effective to promote language learning, students' feedback to me on the materials and the test was important for me. Thus, students were asked to fill out a questionnaire after they finished the post-test.

Did you consider having your students take your test on a computer? Why or why not? If not, what were your concerns?
I did not consider that. The test was designed to be given in a classroom setting and there was only one computer for teacher's use in the classroom. My intuition was to give learners a written test which I thought would be time-efficient for test administration.

Did you consider making the test an open book test?
No. Since the purpose of the test was to investigate how much students can recall what they learned from the lessons and whether the authentic materials promoted their learning, the test was a closed-book test.

What sources did you draw from for your test items?
Test item ideas or content from course objectives?
Yes, I referred to content from the course objectives.
Test item ideas or content from what learners do in class?
Yes, I referred to what learners do in class.
Test item ideas or content from worksheets you make?
Yes, I referred to my lesson plans and handouts of the courses to consider what were the course objectives and further decided what should I include in the test items.

Did you consider learners' communicative competence when writing test items? What aspects of communicative competence?
Yes. Grammatical, discourse (textual), and sociolinguistic competence. For textual competence, I wanted learners to show they could synthesize grammatical knowledge they know (e.g., vocabulary, grammar) and express a textual message in a short paragraph based on what they observed from the film clip (global knowledge: who/what/when/where/how).

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Other sources not named here?
I mainly followed the principles of designing a test from the book that I studied for a graduate course in testing.

Which test item formats do you prefer to use?
I selected the item formats for different purposes. If I only wanted to check learners' comprehension, I will use matching, cloze or True/False. If I wanted to check whether learners know how to use the second language, I will use short response or task items.

What types of learner knowledge do you believe you are capturing in your test? How does that change with test item types you used on the test?
I think I captured learner knowledge on Chinese food culture, vocabulary related to Chinese cooking method, and also learners' ability to connect ideas in a short paragraph. To elicit learners' opinions on Chinese food culture, I used short response items. To check whether learners could recall the vocabulary, I used matching items and cloze. To probe learners' discourse (textual) competence, I designed a writing task.

What learner skills do you believe you are capturing in your test? How does that change with test item types you used on the test?
The test domain is Chinese food culture, which encompasses the constructs including: 1.) Vocabulary related to cooking method in Chinese culture (matching items); 2.) Learners' bottom-up listening processing skill (cloze); 3.) The ability to interpret and describe the major events and characters observed from the film clip (short responses); 3.) To compare and contrast the cultural differences between American and Chinese culture (short responses); and 4.) To develop learners' textual competence, which enables them to develop a short discourse in a cohesive and coherent written paragraph in the target language (either Pinyin or Chinese characters) based on their general comprehension of major characters and events (task).

If you wrote a performance test (where learners have to converse or present a topic, or write something above the sentence level), how did you get the ideas for what task learners had to do for the test?
I got ideas from working with my teaching coordinator. She often designs tasks for students with voiceovers on videos. Since the authentic materials I used was a silent (no speech) film trailer, I thought it would be a great task for learners to write a script for the film trailer.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

How did you get ideas on how to score learners' performances (the scoring criteria)?
From a testing book I was studying. Since students were only asked to compose a paragraph, there were not many constructs to capture. Thus, a holistic scale is sufficient to judge learners' written performance at a specific level. See Table 4-1.

If you used points on a scale for scoring learners' performances, how did you decide on how many points on a scale to use?
I consulted my instructor in the testing course. She suggested using a four-point scale to assess learners' writing skill.

Were you concerned about whether you could get a colleague to help you score learners' task performances?
Not really. There were only five test-takers in the second stage (the pre- and post-test group), so it did not cause much of a burden to grade. I used intra-rater reliability, where I rated the writing task on one day and evaluated them again at a later date.

Did you make another, perhaps simpler or shorter, version of the scoring criteria for learners to use?
No. the scoring criteria was only for teacher assessment of learners.

Did you seek help from a peer to clarify what your test items or tasks were measuring?
Consider: Did you state at any point what you wanted to measure in your test?
Yes, I did a peer evaluation on my test items with my testing course instructor.

Did you make any changes to your test or items as a result of your colleague's feedback?
Yes. Major changes were made in three aspects based on the feedback given by my testing course instructor. First, the test instructions were modified to be more explicit and precise for students to answer. Second, the number and the choices of test items were changed to be more functional. Third, the listening audio file was edited to reduce distracting background noise.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you compare your test to the lessons that learners had?
Yes, I tried to make sure what was tested in the test, students had learned in the authentic materials module.

Section Two: Test Administering and Scoring

Were you concerned about test security? What did you do to ensure test security?
Not really but I made sure the tests were securely placed in my locked office drawer.

Were your tests photocopied, or did learners see your test items or test tasks another way, such as on the blackboard?
They were photocopied.

How did you deal with learners who missed the test, or who were late for the test?
There were two students who missed their post-tests. Since it was important to administer the post-tests right after the instructional module ended, I basically excluded both their pre-test and post-test scores for the purposes of my project.

How did you prepare learners to take the test? Were there any test items or test procedures that were new to them?
I prepared them for the test through the classroom activities they had already had. Similar items showed up when they took the test.

Did you pilot your test? Do a trial run? Did the pilot result in any changes to the final version of your test?
Yes, I piloted the test in Spring, 2018 using a master and non-master group. I revised the test in terms of test instructions, number and functionality of test items, and the listening audio file for Cloze subtest. Then I taught the authentic materials module and administered the test again with a new group of beginning learners in Fall, 2018. For that administration, I used a pre-test and post-test design to evaluate whether students learned the materials and improved on their performance after completion of the courses.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you write any of the test in the learners' first language? Why?
Yes, I wrote English to give test instructions so that students were clear what they were supposed to do.

For classroom tests: How did you accomplish scoring learners' tests?
Consider: Did you score learners' tests twice for accuracy?
Yes.
Did you consider alternate answers and add them to the key?
Yes, I considered the alternative answers in my mind and added them to the key.

Did you randomize learners' tests for scoring, thus erasing any order from where learners sat, or at what point they handed their test in?
Yes.

Did you go back and change your marks on previously scored tests in response to problems you found while scoring tests later in the process?
Yes.

For performance tests: How did you accomplish scoring learners' performances?
Did you go back over your scoring a second time, after the test was over?
Yes.

Do you think your test was reliable? What did you do to check?
Consider: Did you ask another teacher to look over your test before you administered it? What did he or she say? Did you make any changes to your test as a result?
Yes, I showed the test to my testing course instructor to check if my test captured the constructs that I intended to measure. Also, she modified my wording in the test.

For a classroom test, did you use item analyses on a spreadsheet such as item facility, or B-index?
Yes, I used item facility and difference index (DI) (Gorsuch & Griffee, 2018, p. 3 and p. 49). I found that most items discriminated between students' performances on the pre- and post-tests well. 13 out of 17 CRT items' DI were .20 or higher. As to the written performance test task (Part D), while students written score were low in the pre-test, they were

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

average or high in the post-test, suggesting that students learned the materials.

For a classroom test, did you use any statistical reliability estimates, such as Cronbach's alpha?
Not really because the sample size was small (five learners), and a normal distribution could not be assumed.

Reporting scores

Did you report the scores to learners?
No. It was an ungraded test.

Section Three: Using Test Scores

Cut scores

How did you decide which scores were passing or failing scores (cut scores)? How did you decide which scores meant a specific grade on a test?
The focus of the test was to evaluate whether learners learn from the authentic materials. Based on the descriptive statistics for the CRT test with a total score of 17, the mean, mode, and medium were 5.8, 4 & 6, and 6 in the pre-test. The result of the post-test reveals a significant growth on learners' grades by increase of 5 to 6 points with the mean, mode, and medium pointing at 11.6, 11, and 11, respectively. According to the statistics, it is shown that the authentic materials is effective in helping students to learn Chinese language.

For a performance test, did you use your scoring criteria to determine cut scores?
Since this is an ungraded test, it is not necessary to determine cut scores. Based on the scoring criteria with four points on each scale, it is shown that the students' mean, mode, and medium of the performance test increased by one point in the pre-test and post-test, indicating that students learned from the materials and were more comfortable using the target language in the assigned task.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

How did you use learners' scores from this test? Consider: Were the scores for your use only?
Yes, they were mainly used for my research purpose to investigate whether authentic materials are effective to promote language learning.

What was the role of the test score in determining learners' grades?
It was a non-graded test thus it did not affect learners' grades.

Using test scores

Do you feel you learned what you needed to learn from the test results about what learners could and could not do, or what they knew or did not know?
Yes. I learned that vocabulary needed more practice in class.

Did learners' test scores change your teaching?
Maybe in future if I can teach the materials again, I will emphasize more on the parts such as learning vocabulary and listening to native speakers' speaking, which students felt were more challenging, according to their post-test questionnaire responses.

If you could turn back time, what would you change about your test? What would you change about your test administration?
For the second subtest Cloze, I hope to decrease the background noise of the video clip so that students can better hear the dialogue.

Consider: Did learners give you feedback on the test? Did they think the test was fair, or helpful?
Yes, students filled out a questionnaire I gave them about the test and the course. In general, the learners pointed out that the authentic materials motivated them to learn the Chinese language and culture. In addition, learners commented on the test as follows: 1). Spending more time on reviewing vocabulary on the test; 2.) Deconstructing and listening to the authentic, native speaker dialogue more times in class, and 3.) Contending with poor sound quality of the movie trailer.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Section Four: Evaluating and Reviewing your Answers

To what extent is your test here an innovation, or something new, for you?
It was my first time to design a questionnaire for learners to elicit their responses on their learning process, which was new to me. Also, it was new for me to design a test based on authentic materials and to see how learners reacted to the test items, which further shed light on my future teaching.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

A Vocabulary Quiz for ESL Learners at an Intensive English Language School

Dale T. Griffee
ELS Language Centers

Introduction

I was born in Niles, Michigan. After an enlistment in the United States Air Force which left me in Waco, Texas, I attended Baylor University where I studied English and philosophy. Since that time, I have tried to forget philosophy and escape its clutches, but I have never succeeded. I began teaching in 1976 in Japan until 1999 and in the United States until 2015. While teaching in Japan, when I became a tenured faculty member at a Japanese private university, I was required to create and administer achievement tests. I began to understand a major problem of teacher-made tests: Any teacher can make a test so hard that no student can pass it and any teacher can make a test so easy that no student can fail it and either situation can ignore the issue of what students actually learned.

At the same time, I was taking classes at Temple University, Japan and eventually enrolled in their doctoral program. There I took a course in testing with J.D. Brown that radically changed my attitude toward testing. I not only saw that testing was an art and a science, but at heart it was a way of processing information. I saw that the process of creating a test was the same process as creating any research instrument and thus tests had implications for teaching and research. Currently I have authored and published the second edition of *An Introduction to Second Language Research Methods: Design and Data* (Griffee, 2018) . I am obsessed with teachers as researchers, instrument validation, and the research process from beginning (Types of research) to end (Research design and data analysis). Who knows where this is going to end?

The school where the quiz was administered is a private, for-profit international chain with an intensive curriculum including ten proficiency levels from "102" to "112." When a student reaches the final level, the school claims that they should be able to enroll in freshman classes in a

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

university and do well. Classes are taught from Monday through Thursday from 8:00 AM to 5 PM and also on Friday morning. The most commonly taught classes are vocabulary, computer lab, grammar, and reading/writing. Teachers are generally required to have a M.A. in ESL (English as a Second Language) or Linguistics and some teaching experience. At the time of this research, the school was closing this branch due to the lack of students. The school academic director had left, and that position was being filled by a competent newly promoted assistant director.

There were four students enrolled in the vocabulary class for which this quiz was developed. The course met five days a week for 50 minutes each morning. Three of the students were at the 112 level (the highest) and one was at the 111 level. Three of the students were female and one was male. Students were from Taiwan, Saudi Arabia, and Cameroon, representing a variety of L1s and educational backgrounds. All students were planning to enter American universities at the undergraduate level.

Materials and instruction. The required textbook was *Essential academic vocabulary: Mastering the complete academic word list* (Huntley, 2006). The purpose of the book was to study the Academic Word List (AWL) developed by Averil Coxhead (Coxhead, 2000). The AWL is well-known in higher education preparation, and it assumes students know the first 2,000 basic words. The AWL is composed of words that are used by many academic disciplines but are not associated with a specific area of study. For example, the term "communicative competence" would not be in the AWL because it is particular to Applied Linguistics, but the terms "abstract" and "theoretical" are on the list. *Essential academic vocabulary* (EAV) is divided into 20 chapters with names such as economics, psychology, history, environmental science, chemistry and information science. During the time covered by the quiz presented here, we studied Chapter 12 which focused on the content area of history because they had studied earlier chapters in previous vocabulary courses. According to the author, the textbook activities are based on *Eleven principles for designing a vocabulary curriculum* by Schmitt and Schmitt (1995). The chapter was ten pages long and included a reading passage on the Industrial Revolution in England. The learners were asked to contextualize 38 vocabulary items from the passage. There were also twelve exercises ranging from multiple choice items, to short-answer writing, to role plays.

The chapter selection and instruction are entirely the teacher's choice, and it is not expected that any teacher can include all the textbook material. I chose to emphasize the reading passage and the section on making inferences, partly because this was the highest-level class and also because students told me that inference was their weak point.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

The Test

Vocabulary Quiz for Chapter 12

Name: _____ Date: _____

Select ten of these vocabulary items we have studied and write one sentence each. Each correct sentence is 5 points. All writing will be graded on grammatical accuracy, vocabulary and spelling, idea development, punctuation, and capitalization.

1. Dynamic
2. Transformation
3. Expansion
4. Disposed
5. Network
6. Perspective
7. Implicit
8. Accommodated
9. Enforce
10. Sustain
11. Enable
12. Ideology

1. _____

_____

2. _____

_____

3. _____

_____

4. _____

_____

5. _____

_____

6. _____

_____

7. _____

_____

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

8. _____

_____

9. _____

_____

10. _____

_____

According to the reading, life expectancy was 40 years in London but only 17 years in the worst neighborhoods of Manchester. What can you infer about Manchester from these facts?

This test was one of many given during the time I was a substitute teacher for the final weeks of the course. It was created in a couple of days, administered, and graded overnight. I coordinated closely with my supervisor, herself appointed recently. The test has two subsections: one, a vocabulary in context section at the sentence level and two, a short writing essay on inferences from a part of the reading. The first part, to write a sentence illustrating the meaning of a vocabulary item, is also graded on grammatical accuracy, spelling, punctuation, and capitalization because the school is strict on these items and uses these criteria for all writing assignments or tests. From a testing viewpoint, however, it muddies the issue because it might be possible to express the meaning of a vocabulary item but be evaluated negatively on other points. This probably accounts for the less-than-high reliability coefficient. For example, the lowest scoring student (72 points out of 100) lost points when defining the word *implicit*: "His principle in life is implicit what he didn't like." On a five-point scale, he lost three points, but in retrospect I cannot say why, exactly. Did he lose points because the meaning of the target word implicit is not clear or did he

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

lose points because I felt the grammar was inadequate? And if so, how could I tell the difference between lack of clarity in defining the target word and confusing grammar?

For Subtest 2, I graded the short, written answer which constituted part two of the test using four rubrics: Ease of reading, clarity of grammar, inclusion of details and explanation, and clarity of the distinction between the cities of London and Manchester. I applied these rubrics holistically, meaning that I read the answer and made a decision. I did not write out acceptable answers as a guide.

Contributor's Questionnaire Responses

Section One: Test Planning and Writing

Why did you write the test? What were the purposes of the test?
I had three main reasons. First, my school uses the term "quiz" instead of the term "test." So, with that in mind, one of the main purposes of the test presented here was that my school required a quiz to help arrive at students' final grades. I had no choice in the matter. Second, I wanted to gather feedback on the chapter we had just studied. If students had all failed, I would have to think about a review of the material. Third, I wanted to see how they could use their inferencing skills.

How did you decide how many subtests to write?
The number of subtests was irrelevant to me. What was important to me was that the total score come to 100 points. Having said this, the test has two subtests. Subtest 1 is on vocabulary and subtest 2 is on student ability to make inferences.

Did you write as many items or test tasks as you needed, or did you write more than you needed, then pare the number down?
I wrote more items than I needed. I selected twelve vocabulary items, listed all of them on the test, and told students to select only ten to answer.

How did you decide how many test items to write in total?
The test has two subtests. Subtest 1 was close-ended and I created ten items, so I could easily set the score (at five points per item) at 50 points. Subtest 2 was one open-ended item that I could set at 50 points. I wanted the scoring to come out to a possible total of 100 points for easy interpretation. My limiting the items students responded to eleven items, I also ensured the students would be able to finish the test in the fifty-minute class period.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Were you concerned at how long the test would take to administer?
Yes. The class is only 50 minutes and the test had to be administered in that time.

Were you concerned at how long the test would take to score?
No, because I would be grading at home.

Were you concerned how you might use the test items themselves for learner feedback?
Yes, because I believe every test should be a learning experience. On the other hand, I experience going over a test in class with students as stressful, because they then use the occasion to argue to increase their points rather than to learn.

Did you consider having your students take your test on a computer? Why or why not?
No. While the school has a computer lab with more than enough computers for students to take tests, the school also has a policy against students taking classroom tests on computers because of a fear that students might cheat.

Did you consider allowing learners to use additional sources such as dictionaries, or their notes, while taking the test?
No. I personally would not have an objection to using additional sources because in real life we can use them, but the school is very strict against this sort of thing.

What sources did you draw from for your test items?
Subtest 1, the vocabulary subtest, came entirely from the assigned chapter in the textbook we were assigned. Subtest 1, writing an inference statement, came from the students' concerns voiced to me in class.

Did you consider learners' communicative competence when writing test items?
Yes, communicative competence was a primary consideration. The students are used to multiple-choice items, but for my vocabulary test items I asked them to write a sentence using the vocabulary item. That would show me that they could use the word and thus understood the word in a manner that suggested they could use it communicatively.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Which test item formats do you prefer to use?
In subtest 1 I asked students to write a sentence using a vocabulary word we studied in the chapter. Students expected and wanted multiple choice format, but I wanted to test student ability to use the word in context as a way of determining the extent to which students understood the word. In subtest 2, I gave them a passage from the text chapter and asked them what they could infer. I limited their answer for grading ease. These students have the ability to spin out wordy answers.

What learner skills do you believe you are capturing in your test?
Learners' ability to read vocabulary in context, and to make inferences.

If you wrote a performance test (where learners have to converse or present a topic, or write something above the sentence level), how did you get the ideas for what task learners had to do for the test?
Part two is a performance test. I got the idea from two sources. The first source is that the textbook included a long essay that students were required to read. The second source is that early in the course I asked students what their problem was, what problems did they have, and their agreed upon answer was that they had trouble with inference. So, we practiced reading and making inferences. The actual test item is not only taken directly from the textbook, but it was discussed in class. The item was designed so that if a student understood the class discussion, they could easily score 100% on the test.

How did you get ideas on how to score learners' performances (the scoring criteria)?
I don't really know. In general, I like to make tests to 100 points because it makes it easier for students and administrators to interpret the results.

Were you concerned about whether you could get a colleague to help you score learners' task performances?
No. This was an ordinary classroom test which was only one of many grading points. Also, teachers in this school rarely, if ever, are called upon to help other teachers in test scoring. In other words, even if I asked, it is not likely anyone would respond.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you seek help from a peer to clarify what your test items or tasks were measuring?
No, this is not done in my school. And even if I sought help, my peers would not know what I was asking, nor would they be disposed to inquire.

Did you compare your test to the lessons that learners had?
Yes, it was important to me that this CRT (Criterion-referenced test) have content validity. We spent considerable class time on stating inferences from this passage. In fact, we studied the exact same question on the test. If I had it to do again, I would select another passage that we had not studied, but I was pressed for time. I worried that part two of the test was a memory test rather than an inference test.

Did you compare your test to the textbook or other materials learners used?
On the other hand, the textbook did not have specific instructions or a model of inference nor did I present one in class because I did not have one. I could not then and I cannot now explain the mechanics of what an inference is or how to make one.

Section Two: Test Administering and Scoring

Were you concerned about test security?
No, I was not worried about test security because this was a regular class test. and not a high-stakes exit test. Also, I made the test only a day or so before its administration, so there was not much chance of students seeing it.

Were your tests photocopied, or did learners see your test items or test tasks another way, such as on the blackboard?
The test was photocopied at school the day of the test.

How did you prepare learners to take the test? Were there any test items or test procedures that were new to them?
The test reflected class activities. No item or procedure was new to students.

Did you pilot your test?
No pilot. I would not normally pilot a test unless I was developing the test as a research data collection instrument. I was running very much on a day-by-day schedule in this intensive English program, and would not have had time for a pilot. Piloting the test would have ranged from difficult to

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

impossible to accomplish. The class activities, which were very similar to the test, served that function.

Did you write any of the test in the learners' first language? Why?
No, because it would have been against school policy and also I could not have, because multiple students spoke multiple first languages.

Was your test administered on a computer?
No, the school has a strict policy against using computers to take tests, even though the school has a language lab with more than enough computers for each student to use one with several spaces between students. The rationale the school advances is that they are afraid students will use on-line applications to "cheat" on the test. I object to this policy because when we normally write, we can use computers if we wish, and normally have dictionaries and other books and even applications we can use. But the school administration has a policy, bordering on phobic, that insists that any work produced by a student be his/hers and only his/hers. I think it has to do with sales and the claims salespersons made to clients who might want to attend the school, or sponsor others to attend the school.

For classroom tests: How did you accomplish scoring learners' tests? Consider: Did you score learners' tests twice for accuracy?
Yes, I scored students twice.

Did you go back and change your marks on previously scored tests in response to problems you found while scoring tests later in the process?
Yes.

Did you put learners' responses to items into a spreadsheet for further analysis?
Yes, and items were revealed as functioning well.

Did you ask the students themselves to score their own test?
No, because there was no time in an intensive program like this.

Do you think your test was reliable? What did you do to check?
I asked the acting Academic Director to look at my test. She did, and made no suggestions.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

For a classroom test, did you use any statistical reliability estimates, such as Cronbach's alpha?
Yes, I calculated Cronbach's alpha (.79) and KR-21 (.78) on the ten items in subtest 1. I was able to calculate reliability because I had the formulas from Brown (2005) in template form on a spreadsheet, and all I had to do was input the scores. If I did not have those templates, I would not have calculated reliability because of the amount of time it would take. I did not verify the assumptions of alpha such as normal distribution and item equivalence, but I am encouraged by the similarity of the KR-21 coefficient. Given the small sample size (four) and the relatively short test (ten items), I interpret the reliability coefficients as minimally adequate for a classroom test. I think the main source of unreliability for part one of the test is that the construct is not unidimensional. I was grading on the clarity of the definition (the main construct of interest), but also on other constructs (spelling, punctuation, grammar) because they are important to the school.

For performance tests, did you do any rater training?
Subtest 2 could be termed a performance test in that learners had to write creatively at the sentence level and beyond, but no there was no rater training. It was a short classroom test. According to my notes, I graded the short, written answer which constituted part two of the test using four rubrics: ease of reading, clarity of grammar, inclusion of details and explanation, and clarity of the distinction between the cities of London and Manchester. I applied these rubrics holistically, meaning that I read the answer and made a decision. I did not write out acceptable answers as a guide.

Reporting scores

Did you report the scores to learners? If so, how did you report test scores to learners?
Yes, I returned the test to students with my score. I gave them time to look at their test.

What was your goal in reporting the test scores to learners?
I consider the test as a feedback instrument. I think students (and teachers) can and should learned from test items judged correct and incorrect.

Did you teach learners how to interpret their test scores?
No, but I think this is an area I should deal with.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you report the scores to anyone else?
Not directly. The final grade give to the school for accountability purposes includes several areas (attendance, class participation scores) including test scores. Generally, unless there is a student complaint, the school Academic Director (AD) does not review this composite. This lack of review is mainly because grades are coming in from all teachers for every student in the school and have to be recorded in one or two days. The AD spends most of his/her time dealing with students judged to have failed and does not have time to review the grades of those who pass. Learners who fail cannot proceed to the next level in the program.

Did you spend time explaining scores, or answering learners' questions about scores in or out of class?
Yes, a major portion of one class was spent going over the test results. This is a sensitive political area, as well as a sensitive academic area. Students usually argue for a change in grade if they can find any point to contest. Teachers in this school generally find this time stressful and want to get through it as soon as possible. For that reason, teachers prefer test items which can be graded objectively, for example multiple choice formats. Teachers in this school find time spent arguing with students to be not only stressful, but a waste of instructional time.

How quickly did you report scores to learners? Was speed a priority?
Scores are usually reported within one or two instructional days of the test, so I would say they are reported quickly. Speed is a priority because this school advertises itself as having an intensive curriculum. There is pressure on the teachers to pass students. This often results in students at the final, highest level who have been pushed through and as a result they have trouble graduating.

Section Three: Using Test Scores.

Cut scores

How did you decide which scores were passing or failing scores (cut scores)? How did you decide which scores meant a specific grade on a test?
I set my test so that a total would be 100 points. This allowed me to use a traditional understanding of 90 = A, 80 = B, and so on. I felt that if I did not use this standard, it would make score interpretation difficult for students to understand.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

How did you use learners' scores from this test?
Scores were used to calculate part of student's grades.

Did anyone else use the test scores?
The score package was submitted to the administration.

Did learners' test scores have any positive or negative consequences for you, in terms of your institution?
If any students failed because of scores from this test, there would have been serious consequences, especially because the school was closing, and students would not have had any chance to continue taking classes.

What was the role of the test score in determining learners' grades? How much weight did you give your test? How did you decide?
The scores from this test were about 20% of final grades. This was enough to make a difference in that it is unlikely that students could have passed the course if they failed the test, but the scores from this test were not solely determinative.

Were other measures used to decide learners' grades, besides your test?
Class participation, class presentations, attendance, and scores from other tests.

What was the relationship of the other measures to your test?
One score among several others.

Did your test capture some knowledge, skill, or ability the other measures did not capture?
The test was a criterion-referenced test (CRT) and thus scores were designed to reflect knowledge of a specific chapter as well as the skill of inference of reading material. Class participation, students' presentations, and class attendance would not necessarily capture learners' ability to infer ideas from what they were reading.

Reporting scores

How did you report scores to learners? Was timeliness of concern to you?
We spent 30 or 40 minutes of one class period discussing scores. Timeliness was important in that we are (were) an intensive program and had only one or two classes in which this discussion could take place.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you hand the test back to learners? Did the learners get to keep the tests? Or did you take the tests back?
I handed back the tests and normally I let students keep them, but this time I asked for them back because I wanted to keep them on file because of this test project.

Did you teach learners to interpret their scores?
No, instead I put scores on a one to one hundred scale which is the conventional scale. I did this because students and administration have learned to accept this scale as "normal."

Using test scores

Do you feel you learned what you needed to learn from the test results about what learners could and could not do, or what they knew or did not know?
Yes, but it surprised me in that one student performed much better than I expected based on the students' lackluster classroom participation, while another student performed more poorly than I expected.

Did your test change how learners studied?
I have no data on washback. I was brought in as a substitute teacher for the last two weeks of a four-week semester as the school was closing.

Did you spend time going over the test in class?
Consider: Did you mention trouble points as general comments? Did you go over each item?
Did you go over specific subtests? Did learners offer answers? Was going over the test a classroom activity?
Yes, to all of the questions.

If you could turn back time, what would you change about your test? What would you change about your test administration?
I would make several similar tests and use them for classroom practice.

Did others (parents or administrators or colleagues) give you feedback on the test?
No, parents are not involved in the school because they are usually not in the country. The administrator (the acting Academic Director) helped me create the test and approved the final version.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Section Four: Evaluating and reviewing your answers

To what extent do you think you've described recurrent patterns in your work with tests?

A possible pattern could be my view that tests should be communicative. The vocabulary in my vocabulary test was selected on the basis of use, meaning that students must demonstrate they can use the word. Students were not used to this and wanted a multiple choice format.

To what extent is your test here an innovation, or something new, for you?

1.) The line between test items and classroom practice is blurred. Variations of the test could be used for classroom and homework practice. 2.) The test content could be derived from student concern. I tested the ability of students to make inferences from statements based on reading content as a direct result of a class conversation in which I asked students what they needed to work on. They all said that making inferences was something they needed to work on.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

A Simple Speaking Test for an English-Language University Communication Class

Myles Grogan
Kansai University

Introduction

As a teacher of non-English majors in a private university in Japan, I often meet students who may not have a specific reason to learn English, but have to take classes anyway. In the broader Japanese context, most students graduate high school at the A1 CEFR level (The Japan Times, 2018), and their experience of language classes tends to revolve around readings, multiple-choice quizzes, and translation (often at the word level). My teaching has evolved around taking learners' largely passive knowledge, and turning it into a practical ability to communicate with peers. The simple speaking test introduced here has been helping me work towards this mission for the last ten years of my teaching practice.

Much of the literature in language testing and assessment is based on standards, benchmarks, and a psychometric approach, but these have limited applications to my classroom setting. In contrast to the high-school system in Japan, universities seem to be increasingly moving to norm-referenced assessment as part of their accreditation approaches. Teachers may be given guidance as to how many points to assign for each grade level, but they often bear sole responsibility for creating the method of assigning those points. In my current setting, grades are awarded by class on a quota system, meaning that some students have to get the highest available grade and some must get the lowest grade. This is a response to a student body entering the system with an unpredictable level of English. Classroom-based assessment must not only measure the performance of these students, however. It must also help to improve it. Consequently, assessment needs to provide summative information that can be used for a grade, and enough formative information

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

to direct future classes. The test presented here has been my solution to this conundrum.

In essence, the test was designed to encourage interaction on an equitable basis, rather than an equal basis. The test is adaptable to different levels by making simple rule-changes. Students have different language strengths and weaknesses, and class members soon learn to collaborate for a positive result. Simple elements, such as asking groups what they are hoping will happen in the test, help to provide rapport between teacher and student. In addition to an individual score, the group score gives students a simple metric with which to gauge their progress, and most students see a rapid gain over a semester. The result is that, come the time for course feedback, many students say they enjoy the test!

The Test

This speaking test was developed to counter a lack of communication experience in English learners in compulsory university EFL classes in Japan. It combines a group score for *Interaction* with an individual *Skill* score to reflect classwork and communication. It is part of summative assessment, but also provides formative data by identifying necessary areas for more practice or identifying student needs.

Students take this test twice during a course, and they have two topics in class before each test. Although general topics come from prescribed textbooks, the specific tests reflect student experience of the topics as far as possible. Recent examples for higher-level learners have included the students' experiences of gender roles or written literature. Lower-level students have talked about part-time jobs or travel. The test is based around a four- or five-minute conversation on one of the two classroom topics preceding the test, followed by immediate feedback from the teacher. Students are assessed in groups of four or five people, which seems to be the ideal size in this setting. This format is regularly used with the students in each weekly class as a self-graded concluding activity. Assessment and classwork are thus directly connected.

Students receive two scores. Their *Interaction* score (out of 10) is based on the number of utterances each member makes. Each member gets the same *Interaction* score, so team members must collaborate to ensure everyone is communicating verbally. A big problem is students who talk too much or too little. As either of these is problematic in terms of a group *Interaction* score, there is positive washback in terms of test performance. The collaboration required for a good score also ensures a positive class attitude once students see the underlying approach needed for the test. A

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

simple image which seems to work is that many students see speaking as *bowling*. They give an idea and withdraw until the next turn. This test, however, requires the passing and teamwork of *volleyball* to be successful.

The *Skills* score (out of 10) is a qualitative measure of each member's performance, reflecting the way students use the skills and techniques covered in class. It may reflect good use or attempts at using language from the course (grammar or vocabulary). The score can also reflect a student's use of communication strategies, such as asking the same question twice (known as a "Double Question" in class) or using paraphrasing. A range of conversational skills can be introduced and used as part of the class to help students gain points in the test. See Table 4-2.

Table 4-2: ***Individual skills scoring rubric***

6 points is considered a passing grade. Slight changes can be made to match levels of classes or individuals if it is equitable to do so. The main goal is that students actively support each other and use different strategies (given the level) in order to make communication a success.	
Score and description	**Conversation features**
0 -2 points: People scoring 2 points or less contribute very little to a conversation. If a member scores in this range, the team score is usually low. It usually means the team has allowed another member to speak too much or too little. Scores at this level indicate a fundamental problem. It may be that a teacher should intervene in the test to keep things equitable.	Learners at this level: ● Show minimal participation in conversation ● Make almost no attempt to speak ● Answer only to direct questions ● Use minimal strategies ● Respond very slowly
3-4 points: People scoring 4 points or less have generally not actively participated in the conversation, but may have attempted short bursts. A grade of 4 or less may also occur when other speakers do not allow members to participate, or they dominate the conversation.	Learners at this level: ● Usually use short turns (one-word answers or fragments) ● Try few strategies ● Do not seem to monitor conversation or other speakers ● Have long silences with slow responses ● Speak with limited focus ● Respond slowly

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

5-6 points: A score of 6 indicates a basic level of participation, with a small range of skills used, but some limited success in communication. Typically, a conversation at this level sees every member scoring, but there are occasional breakdowns, or speaking remains limited in depth.	Learners at this level: ● May move off-topic ● Show limited response/follow-up and superficial focus on topic ● Often focus on particular speakers and excludes others ● Display long silence and slow responses ● Show limited, often short turns
7-8 points: A score of 7 or 8 indicates a reasonable variety of strategies, and a good attempt to pursue the topic. Although some Japanese (or L1) mannerisms may appear, they do not intrude on the conversation. Breakdowns may occur, but a speaker makes a good attempt to sustain the conversation with all speakers.	Learners at this level: ● Use a mix of strategies ● Have both long and short turns ● Are "noisy" listeners ● Ask follow-up questions ● Uses some vocabulary, grammar, and language skills discussed in class ● Show a positive focus on using English ● Try to monitor and respond to other members
9-10 points: A speaker scoring 9 or 10 has made a significant contribution to the group. Although language may not be perfect, the speaker's contribution to the conversation is usually positive and enjoyable.	Learners at this level: ● Stay focused on the main question or related topics ● Include all other speakers ● Give extra information (spontaneously) ● Use English thinking and conversation sounds ● Respond quickly (e.g. echo key words, repeat question) ● Use vocabulary, grammar, and language skills discussed in class ● Mix long and short turns ● Use multiple communication strategies for self and encouraging others ● Help others to speak

Using this test process, I have found that tests take slightly more than ten minutes per group. I have successfully finished eight groups of four or five in one 90-minute class. It is worth practicing the procedure in class (for example, the week before) to make sure that everything goes smoothly. The scoring and format are highly adaptable, and various alterations can be made

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

to allow simultaneous interviews, or cope with other necessities of the classroom situation.

Classroom lead up. Following a basic introduction to classroom communication in the first two lessons, students are introduced to the first of two topics. Topics from textbooks are distilled into a simple discussion question that allows students to share personal perspectives and opinions. Example textbook topics, like "Human Migration" (Table 4-2, Topic 1), become centred around a main speaking practice question, which is also used as the focus question for the test. Practical language skills to help interaction may then be introduced progressively. One of the two questions is selected for a team to display these skills in the test, concluding that section of the course. This means that students should prepare for both questions. An example course plan leading up to the test is shown in Table 4-3.

Table 4-3: ***Course plan leading up to the test***

Class	Focus	Content
1-2	Introduction to course goals	Basic communication phrases and skills
3	Topic 1: Human Migration Input Introduce test question 1 : *"Who do you know that has relocated to another place? Tell their story and comment on whether you could do the same."*	Simple thematic group Q & A, basic vocabulary, listening, first try at speaking topic HW: Simple essay
4	Topic 1: Output Practice test question 1	Practice test, language skills, guided peer review of essays in discussion format, matters arising from class, final pass at test question
5	Topic 2: The Natural World Input Introduce test question 2: *"Describe a natural place that you like (such as a National Park). What is it that makes it special?"*	(As week 3)
6	Topic 2: Output Practice test question 2	(As week 4)
7	Test practice	Review of skills practiced
8	Test	

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Test process. Students are formed into groups of four or five people. This can be done before class or on the day of the test, depending on reasonable norms for the school setting. Students in a team are invited to be seated wherever they like around the testing table. As a warm up, the teacher may review points the students should be careful of, and ask some simple questions to get people speaking. A good example may be which topic the students would prefer and why. Following this, a coin is flipped to choose the test question. Tests may be video-recorded for further review and refinement of the test process or to help identify formative points. If this is being done, the video may be started after the warm up. The teacher must then fill out the Grade Recording Sheet (Figure 4-1). Students introduce themselves again to the teacher, giving their names and student numbers. These are recorded on the Grade Recording Sheet in the grey section. If a video is being used, be sure to get the introductions on video so that points can be assigned properly if a review is conducted in the future.

Date: _____	Class day: _____ Period: _____ Semester: Test #:		
Name ID	Name: ID: _____	Name: (Middle student) ID:	Name: ID: _____
Personal	*Clarification* *Echoing/reactions* *Language/grammar* *Helps team mates* *Conversation strategy*	*Clarification* *Echoing/reactions* *Language/grammar* *Helps team mates* *Conversation strategy*	*Clarification* *Echoing/reactions* *Language/grammar* *Helps team mates* *Conversation strategy*
Group			
Name ID	Name: ID:		Name: ID:
Personal	*Clarification* *Echoing/reactions* *Language/grammar* *Helps team mates* *Conversation strategy*		*Clarification* *Echoing/reactions* *Language/grammar* *Helps team mates* *Conversation strategy*
Group			

***Figure 4-1*: Grade Recording Sheet**

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

In the current setting, experience suggests that one minute per person provides a reasonable time frame for the conversation. Hence, four people will talk for four minutes, and five people will talk for five minutes. While people are speaking, the teacher records the *Interaction* score as a five-bar gate (see below for scoring). The teacher uses the box on the "Personal" row to record *Skills* demonstrated from the class with a check mark. Examples of particular strategies or linguistic forms may be recorded in or near each participant's record. Hence, an indirect question ("*Can you tell me if* you've been to Tokyo?") may be recorded as "I●." Space permitting, recording specific examples is useful for post-test feedback. After the timer sounds, the teacher gives feedback to the students. If possible, tailoring feedback to previous classwork is most useful to the students.

Scoring. The *Interaction* score is based on the number of "active verbs" each student gives. For most students, this means sentences. However, a more complex sentence, like a relative clause, can get two or three points. In the following example, note that some verbs (such as "go") get no points, as they do not carry any additional meaning: They are not "active" verbs.

"I *like to* go shopping" = 1 point.
"When *I'm* free, I *like to* go shopping" = 2 points.
"What I *like to* do when *I'm* free *is* go shopping" = 3 points.

Students can practice using this scoring system, first with the written homework (see lesson 3 or lesson 5), and in classroom conversation. A target is given for particular grades (e.g. C = 9, B=10-11, A=12-13, etc.). Low level students may get points by echoing or conversational shadowing (Murphey, 2001). Higher level groups may be told that repeated phrases will not get points, requiring a wider range of phrases. Teachers may support weaker students by arranging a signal for a memorized chunk, although groups often do this for each other. The teacher may also wish to indicate when a strong speaker has achieved the target number of sentences and should let others speak.

The *Skills* score comes from the use of skills covered in class. This may be basic communication skills, such as asking for clarification. Basic phrases for different functions can be introduced on a *Skills card* (see Table 4-4 in the Contributor's Questionnaire Responses section) and practiced in the first two classes, using appropriate activities (simple card games, gap fills, etc.). Further communication strategies, such as asking the same question in two ways ("Are you hungry? Do you want to eat?"—a "double question"), can be introduced in a unit. Linguistic features covered in class, such as relative clauses or the use of perfect tenses, can also be recorded,

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

allowing teachers to see quickly who has used the skills. A completed example of the Grade Recording Sheet is given below (Figure 4-2).

The actual skills recorded and used for grading will require some adaptation by group. It is helpful to prepare an example rubric, based on the teacher's experience of the class. A simplified rubric can be shared with the group before the test, but given that the goal is to help students fairly, it should remain flexible.

Date: 9th Nov, '18	Class day: Mon Period: 3rd Semester: Fall Test #: 2		
Name ID	Name: Hiroshi Tanaka 8 ID: 2222	Name: (Middle student) ID: _____	Name: Saori Nakajima 10 ID: 3333
Personal DQ IQ	Clarification // Echoing/reactions Language/grammar / Helps team mates Conversation strategy ///	Clarification Echoing/reactions Language/grammar Helps team mates Conversation strategy	Clarification // Para Echoing/reactions // Language/grammar // Adj. Ph Helps team mates /// Mon. Conversation strategy // Exp.
Group	卌 卌 |||		卌 卌 卌 ⊕
Name ID	Name: Hiroyuki Uchiyama ID: 1111 7		Name: Sachi Suzuki 9 ID: 4444
Personal Frag	Clarification / Echoing/reactions // Language/grammar Helps team mates Conversation strategy /	8	Clarification // Para Echoing/reactions // -ve Language/grammar // 10, A.f Helps team mates / Conversation strategy ///
Group	卌 卌		卌 卌

Notes. DQ – Double question. IQ – Indirect question. Para – Paraphrase. Adj – Adjectival clauses. Perf. – Perfect tense. Exp – Expand. -ve – Negative disagreement. Frag – Fragment.

***Figure 4-2*: Completed Grade Recording Sheet**

Contributor's Questionnaire Responses

Section One: Test Planning and Writing

Why did you write the test? What were the purposes of the test?
The test was specifically designed to meet the institutional requirement of a midterm and final for a course. I wanted the test to reflect activities that could be used in class, with identifiable skills from the class situation, and with a suitable way for students to conceptualize their progress. However, my overarching motive was to help students to be more comfortable with English and with interaction. Reducing the affective barrier to communication was a major goal of the test.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

How did you decide how many subtests to write? Was there some correspondence between a subtest and a course objective, or was there some other reason to create the subtests you did?
The test score consists of two components (*Interaction* and *Skills*). The ability to interact struck me as being the main element that was lacking in the classroom and in student development. In addition, I needed more concrete objectives to help the interaction come about, hence the particular *Skills* section.

Were you concerned at how long the test would take to administer?
Administering the test in a specific time was a concern at first. I have 90 minutes per class to complete all tests. I have now done this test four times a year for over eight years, usually with seven to ten classes. On only a couple of occasions have I gone overtime, and even then, not by more than a minute.

Were you concerned at how long the test would take to score?
Scoring based on the counting system is more or less immediate. The skills section takes a little more consideration, as I usually have to consider how the individual has responded to specific class-based coaching (e.g. a quiet student trying to speak more, or a noisier student actively taking more time to listen to classmates).

Were you concerned how you might use the test items themselves for learner feedback?
Given the time constraints, immediate feedback is limited. However, I always try to establish short, individualized coaching points for the student. This just seems like good manners! Doing so in two or three minutes, however, does create time pressure. It has taken practice to know what to look for.

Did you consider having your students take your test on a computer? Why or why not? If not, what were your concerns?
The main issue here is that interaction would be more cumbersome and removed from a normal environment. I have considered asking the students to take the test as an "open" test, so that groups of students just hand me a video file for grading. I think this would be logistically quite difficult. For example, if I could not hear people talking, it would be impossible to grade.

That being said, sometimes students miss the day of their test, owing to ill health or other reasons. For these students, I usually request a video made

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

with classmates. This works quite well. I can ask about any particular issues if the numbers are small. I could not do that with over 200 participants.

Did you consider making the test an open book test?
Getting students to be aware of the difference between scripted episodes and naturally emergent conversation is difficult. Even as it is, students often try to write "scripts" for the conversation. I do not mind staged communicative episodes to show me a skill, but the idea here is to see more spontaneous conversation, rather than reading prepared dialogue. (It is for this spontaneity that I toss a coin to choose the topic at the beginning of the test.)

Did you consider allowing learners to use additional sources such as dictionaries, or their notes, while taking the test?
My graduate school insisted "You are your own best resource," and this is a message I try to encourage students to learn. Although I do not allow external resources (spontaneity again), I actively encourage students to support each other during the testing process, with phrases or strategies and so on. I really hope this carries over into their real-life experiences.

Did you plan to allow learners to re-take a test for improvement? The same test, or a different test?
On the one or two occasions in the past that I have had groups freeze, fail, or otherwise not do what I know they are capable of, I have allowed retakes. This only tends to happen in the first test—it may be that they cannot accept the procedure mentally, as it's quite different from their previous experience of "tests." On these occasions, I have stopped groups and told them to rethink what is happening. After directly addressing whatever issue is happening—say, a single member is not letting others speak, or is not participating—I give the students another chance. An upper limit of an A or a B may be imposed, but students may try again after the remaining groups have finished their test, or first thing in the next class.

What sources did you draw from for your test items?
The first consideration for skills is a needs analysis based on classwork. The first test I give consists mostly of what might be called "Classroom English," and is focused around establishing a base for communication. Most textbooks have a "Classrooms skills" section in the first unit of most textbooks, or at the back in a skills reference section. I transfer these to a "skills card" to make them more tangible. These are supplemented with resources from conferences and in print. See Table 4-4.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

In particular, context-specific pragmatics resources, such as the Japan Association for Language Teaching "Pragtivites" book (Ronald, Rinnert, Boyd, & Knight, 2012), have helped with needs analysis and identifying new skills. Input from such resources can be added to the skills card or used to supplement it. The card itself can be used, in whole or part, to provide practice or reference in class. I place the skills card and any other similar resources the class Learning Management System for review or for anyone who misses class.

Table 4-4: ***Skills card for students***

Asking for opinions	**Clarifying**
• What do you think? • How do you feel about . . . ? • What's your view on. . . ?	• Sorry. What do you mean? • I'm not sure I understand. • Did you say ("*collect*" or "*correct*")?
Giving opinion	**Reacting**
• Oh, I feel . . . • Well, I believe . . . • From my point of view, . .	• That's *strange / fantastic / terrible!* (Comment) • I'm sorry / happy to hear that. (Comment) • I hear / saw / /read that (+time/place)
Disagreeing	**Paraphrasing**
• I don't think so. (Comment) • I'm not sure about that. (Comment) • No way! (Comment)	• You mean . . . • Are you saying . . .? • To put it another way . . .
Agreeing (positive)	**Showing no opinion**
• I think so too. (Comment) • That's right! (Comment) • I couldn't agree more. (Comment)	• I really don't know. (Comment / question) • I haven't given it much thought. (Comment / question) • I can't make up my mind. (Comment / question)

Occasionally, some members may need special consideration, such as those who suffer from anxiety. In these cases, the "source" for assessment may have to be my own observation of what the student can do. The skills on the card may need adapting or changing completely. These people may be expected to show specific behaviours in line with particular needs. One student, for example, wrote a script for a group to show the skills from class.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

What types of learner knowledge do you believe you are capturing in your test? How does that change with test item types you used on the test?
The test's main focus is probably skills and attitude. I encourage students to display certain grammatical features (e.g. indirect questions, real conditionals) from the content covered, as well as vocabulary from the unit or their own investigations as appropriate. My main concern, however, is an overall willingness to participate and willingness to improve.

What learner skills do you believe you are capturing in your test? How does that change with test item types you used on the test?
The majority of the skill is based around building good communication. The focus is pragmatic and social, showing an awareness of other interlocutors. Aside from asking for clarification or communication repair, students may also try asking the same question in two ways ("double questions" allow an interlocutor more context and more time to answer), sharing similar experience (aligning towards the speakers), turn-holding, and making sure that all members have a chance to participate. In addition, we look at ways of communicating with people of a higher or lower language level, as they may need to do that in real life (be it their L1 or L2).

To some extent, the main learning is attitudinal, in that I am looking for students to locate weak points and work on them. For example, where students have a tendency to speak rather than listen, addressing that will get higher points. Where students may have difficulty speaking with others, initiating a turn, or even a verbal acceptance of an utterance helps to raise the grade.

If you wrote a performance test (where learners have to converse or present a topic, or write something above the sentence level), how did you get the ideas for what task learners had to do for the test?
Tasks and topics are taken from compulsory texts and adapted as a conversation topic.

How did you get ideas on how to score learners' performances (the scoring criteria)?
The seed of the idea came from a conversation with a colleague who had seen a conference presentation on a speaking test in which each sentence was given point. I tried this much later as a simple class activity.

The first major practical problem came with the idea of what a sentence was. I found the simplest idea was to identify active clauses (i.e. verbs carrying marking for person or tense/aspect). This is a broadly simple technique which can be explained to students easily. I have had to adapt the

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

scoring slightly for different classes, such as not allowing points for repeated sentences, but overall it seems to work.

The second problem was with students who would either not participate or who dominated the conversation. This was more difficult, but there is a Japanese expression meaning "joint responsibility." I heard the expression among the students a few times, and decided it could be used. This expression is well understood, and it seemed to appeal to the groups.

If you used points on a scale for scoring learners' performances, how did you decide on how many points on a scale to use?

Put simply, the university requires each test be worth 20% of the course grade. Scoring each of the two sections out of ten points makes placing students on the scale simple.

The *Skills* score needs constant updating with each course I use it for, and it can even be negotiated somewhat with individuals with specific needs. It has largely been a matter of experience. I have generally found that writing out my expectations for students at certain intervals helps. I have shared the rubrics that I have made with other teachers at irregular intervals, but regular calibration through reflection is necessary.

Getting the *Interaction* score well-balanced is also a challenge. After students have some experience of the system from the first semester, I usually raise the amount of points required in the following semester. This can work well, by giving students a new challenge. Sometimes, however, students get overly tired of pushing themselves, and maintaining motivation can be hard. In some classes, most of the students get very similar scores, and groups seem to gravitate the B or A level.

The school, however, requires specific quotas of grades. In some classes, this means that between 20 and 30 percent of students must get the highest course grades, while 50 to 70 percent must get the middle grades. Besides measuring performance and promoting learning, the test must rank the students. Creating an interaction score that ranks students fairly, based on their progress in class rather than their ability when they arrived, can be problematic. In these cases where groups of students get similar scores, I tend to believe the problem is mine for not identifying and addressing issues with motivation. In response, I usually try to find an equitable solution, such as identifying other aspects of student performance from the course that may legitimately help balance points awarded to students in a reasonable way. This might be performance in another areas (such as homework completion), or scores based on classroom participation.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Were you concerned about whether you could get a colleague to help you score learners' task performances?
Getting colleagues to score tests in the context may not be impossible, but it is far from practical. I have, however, had other teachers check rubrics, or discussed aspects of how pedagogy and assessment may be made fairer.

Did you plan to give the scoring criteria to the learners for future self- or peer-assessment? Did you make another, perhaps simpler or shorter, version of the scoring criteria for learners to use?
Students are given an outline of the scoring process early on, usually on the third class. Because of the flexibility required for each class, these are not absolute. The students are given plenty of opportunity to identify the skills they will be required to show, and they regularly practice scoring their own performance on the *Interaction* section as a weekly activity.

In the week prior to the test, students are assigned to observe a different group. They grade an individual using the *Interaction* scoring system and comment on (rather than score) observed performance in the *Skills* section. They may also be given a paper-based quiz, such as a discourse completion task, to specifically target skills or scoring.

Did you compare your test to the lessons that learners had?
Skills assessed reflected those practiced in class. These skills are listed on the *Skills* card (Table 4-3), which is made available on the class Learning Management System. Aspects of the lessons are a feature of test feedback.

Did you compare your test to the textbook or other materials learners used?
The class content is built to reflect textbook topics and the opportunities for speaking afforded by it. The test is therefore built around these learning opportunities, and the material covered may be used in the test feedback.

Adapting existing tests

Are you required to use specific tests in your program?
Besides the test mentioned here, there is a common online unit that students must complete. For the program under discussion, it currently represents 10% of the grade. In addition to the test here (given twice, at 20% in each instance), another 30% should come from classwork, and 20% from "small tests," such as weekly vocabulary quizzes that a teacher may develop.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Section Two: Test Administering and Scoring

Were your tests photocopied, or did learners see your test items or test tasks another way, such as on the blackboard?
Prompts are displayed, although they are seldom used. The printed question seems to work as a "way in" for some students, getting them started. It can also be a way to return to the topic if group members begin to stray off topic.

How did you deal with learners who missed the test, or who were late for the test?
Students are usually asked to video the test on their phones. They need to get other people (usually class members) to help them reproduce something close to the actual test, and provide me with a video file of the procedure. On one occasion, a student got help from people not in the class. The non-class participants even used some of the language from the *Skills* card (Table 4-4), so the student must have taught the other participants first!

How did you prepare learners to take the test? Were there any test items or test procedures that were new to them?
The test is based on a lesson cycle. Students will have practiced keeping their own *Interaction* score, and are regularly asked to use the *Skills* in the practice. I give coaching and extra practice in skills where possible. Students have usually done a full-length practice on each topic at least twice before taking the test.

Did you pilot your test?
The test has been under continuous development. It began as part of a project to encourage long-term low-level students to participate. From a classroom activity, it has become a cornerstone of my teaching.

Did you write any of the test in the learners' first language? Why?
In the current setting, I use the first language very sparingly, although I have used more L1-based introductions to the test elsewhere. For many learners, "speaking English" means the ability to translate English into their L1. Translating has a lot of academic cachet, where conversation does not. While recognizing the importance of translation as a skill, my focus is to build the importance of L2 communication and interaction. The test is designed to emphasize the more immediate form of communication; what my first TEFL teacher called a "language bath."

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

For classroom tests: How did you accomplish scoring learners' tests? Consider: Did you score learners' tests twice for accuracy? Did a colleague help you score? Did you later change your scoring because you recognized some learners couldn't answer certain test items due to some aspect of the items or the test?

Tests are videoed, so double scoring is a possibility. Tests are always available for learners to review, but none have asked to do that (which is hardly surprising in the context). It is not really practical, however, to repeat seven 90-minute classes of assessment in a week.

Occasionally, an entire class gets relatively low points in the discussion. In this case, I am willing to make adjustments while reviewing performance. This test only represents one part of the grade, albeit a large one. From the institutional perspective, the goal is to produce ordinal data, where one learner is in first place, another learner is in second place, and so on. Because a certain number of S, A, B, and C grades must be recorded to meet institutional grading requirements, it is important to produce a ranked profile rather than a particular criterion-based measurement.

Did you put learners' responses to items into a spreadsheet for further analysis? Did that process help you catch scoring accuracy problems? Or problems with bias?

Following a test, all results are laid out in a table format for me to reconsider and review as appropriate. Grading is relative, so if it seems that a group performed better than others but that performance is not reflected in a higher score, I will reconsider the grade. I tell my students that I may move some scores up, but I will never move scores down.

Did you ask the students themselves to score their own test? Or a classmate's test?

In the past, I asked students to score for every member of their team on *Interaction* during class practice, and to compare results. This produced some useful discussion, but was very hard work for all involved. Now, I ask students to score only their own practice sessions. This can be distracting, and some speakers may get so excited they lose count or forget completely. Fortunately, students generally remind their colleagues to count.

For performance tests: How did you accomplish scoring learners' performances?

Consider: Did you record learners' spoken performances to score later, or perhaps score a second time? Was recording equipment available? Did you

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

score learners' spoken performances at the same time learners gave their performances? If so, did you have enough time to score?
I use a five-bar gate counting system on the Grade Recording Sheet for the *Interaction* component. I have reviewed samples of these scores in the past, but seldom do so now. I believe I produce a fair grade using this system.

For the *Skill* component, there is a checklist on the Grade Recording Sheet to mark-up particular elements. The skills I am looking for have been mentioned in class often. The pace of the test can prohibit taking detailed notes of each student's language, but I try to note one or two features for use during feedback to discuss with the student in the (brief!) period after the test.

Do you think your test was reliable? What did you do to check?
I am not sure that psychometric reliability, such as may be expected in large-scale, multiple-choice tests, is a reasonable expectation of a classroom test. Many things can affect student performance in interaction, from linguistic skills to physical condition. I do try to take account of factors so that students would achieve a similar score if they were to take the test again.

I would emphasize fairness over psychometric reliability. The score in this test is dependent on the performance of others in the interaction section, and changing the group can have notable consequences (though this effect seems to diminish over time, as students get used to it). My goal as a teacher is to prepare students to interact with a variety of other interlocutors. These interlocutors may be of higher or lower general ability, have different personalities, or have different background knowledge. The factors that the students deal with in test—the attitude of group members, their pragmatic awareness, and even the culture of those in their team—are a reasonable reflection of what may be found outside the classroom. This aspect of communication affects reliability in the psychometric sense, but it is something the students learn about and they can incorporate it into their test behaviour.

I would say that in general my grades are a fair reflection of the performance that a student gives. They also seem reasonably consistent with expectations based on classwork. The more familiar a class gets with the communication skills underlying their linguistic performance, the more reliable (in a general sense) the grade becomes.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Reporting scores

Did you report the scores to learners? If so, how did you report test scores to learners?
I tell the scores to the students before they leave the test room. This feedback is in their L2 (English). I do this in the groups, giving both number grade and feedback, with a short evaluation of their performance. Numeric scores are also posted to individuals via the Learning Management System.

What was your goal in reporting the test scores to learners?
Many students do not appear to keep track of their grades, and may only wish to pass. Part of the goal is to change this behaviour and encourage a more active approach to learning. For me, the score is to help students develop, by identifying concrete goals or areas for further reflection. Possible actions and strategies to improve learning are, in many ways, more important than any single test score. This is part of the reason for immediate post-test feedback.

Did you teach learners how to interpret their test scores?
The main focus of the "interpretation" is the performance assessment, mentioned above. I would be happy to dispense with the numbers and just go to pass/fail, but the grading system at universities seems to be a part of funding requirements. The main aim of the test for me is to use feedback as part of coaching.

Did you report the scores to anyone else?
No. The school simply requires a numeric course grade, of which the results from this test make 40% in total.

Did you report peer-assessment scores or self-assessment scores on the test?
No, but an active approach to self-assessment in the test practice forms part of a different aspect of the overall course grade ("Contribution to class").

How quickly did you report scores to learners? Was speed a priority?
As mentioned above, speed is definitely a priority.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Section Three: Using Test Scores

Cut scores

How did you decide which scores were passing or failing scores (cut scores)? How did you decide which scores meant a specific grade on a test?
The institution has specific percentages relating to each grade. See Table 4-4. Cut scores for both skill and interaction test are related to experience of these percentages. Note that the institutional guidelines are only numeric, and not linguistic. This means there is no reference to an external measure, such as CEFR (Common European Framework of Reference; Council of Europe, 2001, 2018). While standards would be nice, the students do not arrive "standardized" in terms of ability or goal. In fact, the placement test used for the program is done on purely receptive skills, so the willingness to interact with others—which will greatly affect performance—is far from standard. The lack of required attainment goals does not allow comparison on linguistic terms, but it can allow teachers to adapt assessment to the needs of the learner. After all, students still have the opportunity to take large-scale tests at any time.

Table 4-5: ***Institutionally required grading percentages***

Score	Grade
Below 60	D
60-69	C
70-79	B
80-89	A
90-100	S

How did you use learners' scores from this test?
I hope the scores inform students of their progress. As mentioned above, it is part of the overall coaching pattern. No data from the smaller aspects of the course grade, such as individual test scores, go to the school or any other stakeholder except the student. There are no direct positive consequences from a good score on the test itself, although it may be necessary if students need a good overall grade. The class is compulsory, so failure or low scores can have a negative impact. Scores are used for calculating GPA (Grade

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Point Average), and students must receive at least a C in this course to graduate.

What was the role of the test score in determining learners' grades?
The institutional requirements for the class have recently changed, but still reflect a multifaceted grade. There are five subsections to the final grade that are mandated by the university.

20% Midterm (This test)
20% Final (This test)
20% Short class tests/quizzes (Previously 10%)
30% Class-based activity
10% Online learning component (Previously 20%)

This helps to prevents "exam hell," and the over-representation of a single construct or test. I try to avoid having the final test on the last week, because of its large weight. I use a small class test worth about 5% on the last day. This means that I can focus on cases where grades are on the borderlines more concretely, and take steps to rectify any potential problems or unfairness. This may be a simple discussion or a closer inspection of classwork, or it may be some kind of task to ensure that a certain grade is merited.

The online component is common to all classes. For this course, it consists of vocabulary study, based on a levelled quiz attuned to individuals. This is not linked with a textbook, but rather is based on wordlists from corpora (Japan Association of College English Teachers, 2016). The "Quizzes" in my case were vocabulary arising from class, making 10%. With the recent addition of an extra 10%, I have added two discourse completion tasks to promote scores in the speaking test. In this, I ask students to use the skills I will be looking for more explicitly, allowing a review in the week preceding the test. This, however, is very much in an alpha stage at present.

Class-based activity is assessed weekly. Part of it is awarded on the basis of proper preparation, such as submitting notes or completing online preparatory assignments. This is done on a "Completed/not completed" basis. Further points are given in terms of how much students interact with each other in the target language, answer questions, and contribute to the learning of their classmates.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Reporting scores

Did you offer feedback to individual learners in addition to their test scores? Written? Orally? In or out of class?
Feedback is verbal in their L2 during the weekly class. More tailored specific advice seems impractical. There may be a way to do it, but it is beyond my current resources.

Using test scores

Do you feel you learned what you needed to learn from the test results about what learners could and could not do, or what they knew or did not know?
This is ongoing and iterative. I use class to inform assessment and assessment to inform class.

Did learners ask you about the test itself (not the test scores) outside of class? If so, what did they want to talk to you about? Did you use specific subtests or items or tasks to focus your talks with learners?
The biggest concern that students have approached me with is that other people may be speaking too much or not enough, or they may have been concerned about their own speaking. Specific coaching may take place, such as asking more talkative students to give others some space (perhaps with some sign—such as using a "passing gesture"—to invite people to speak). Another example might be helping students craft particular episodes to use in their speaking. This can be done with homework or via some other system.

I try to build rapport with students so they are confident enough to consult me about such issues. As a result, I include the use of specific strategies (or not) that have been negotiated with students in the scoring where appropriate. This is part of the "equitable" aspect of my approach to assessment.

Did learners' test scores change your teaching?
To some extent, it has re-emphasized the focus of my classes onto practical communication and conversation. While academic presentations are important, for example, I see conversation as being a requisite for other skills. Presentations may be viewed as a one-sided conversation. I have re-focused listening activities into note-taking exercises (usually as homework), and now I emphasize re-telling of presentations. This allows for repair, language noticing, clarifying, and other activities in way that simple quizzes don't.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

This test is used specifically with my first-year students. I have made special efforts to steer away from this approach with second-year students, so as not to over-do the approach used here. It is useful, but other work is needed too.

If you could turn back time, what would you change about your test? What would you change about your test administration?
If I were to start again from scratch, I would try to more concretely combine theory and practice. I would try to emphasize the communication strategies more concretely using discourse completion tasks, highlighting the declarative knowledge side of things. This task may fit elsewhere in the classwork or short test section of the score. This would hopefully make the tasks clearer overall.

I have also wondered about asking students to transcribe the test. I do not think this is practical, and it has the potential to be quite boring or overly detailed. In addition, the students I teach seem to have a tendency to script everything, which would potentially turn this into a screenplay event. Some may try to write up the conversation before the test rather than after it.

I would probably think differently about transcription if the students were communication majors. Some form of transcription would be helpful in making the grade transparent, but these students are taking a required class. The usefulness in terms of learning may be negated if the transcription task were too onerous.

Section Four: Evaluating and Reviewing your Answers

Please read through your answers to the items and answer the following:

To what extent do you think you've described recurrent patterns in your work with tests?
I think the teaching cycle and the personal relations between students appear to be a pattern. A concern, however, is the lack of connection to external supervision or meaning attached to the grades, although I believe my situation is close to the norm for this context in that respect. The ultimate responsibility for the consequences of the class and the grading is mine. Although I believe my classes are useful, it is hard to see how they fit into the broader picture with the students' university careers and their lives in general.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

To what extent is your test here an innovation, or something new, for you?
I believe this test is innovative, in that the washback promotes the teamwork and collaboration that is missing in exam-focused classes that seem to be part of high school in Japan (although this appears to be changing). It allows me to let the students talk in class and shows a direct link between talking and "speaking English." Finally, besides the specific skills, it also shows students a concrete measure of their ability to interact, and most students will see a rise of around five points (more in many cases) in their *Interaction* score over the course of an academic year.

There are flaws in the test, and it remains a work in progress. It is hard to find specific external measures that it may be parallel to. My main achievement, however, is that students get to know one another and develop a more positive attitude to communication. Some of them even enjoy the test, which is not a common occurrence in language classrooms.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

An English as a Foreign Language Test for Reading, Writing, and Cultural Diversity Awareness for High School Students

Juliana Jandre
Federal University of Rio de Janiero
& Vander Viana
University of East Anglia

Introduction

This is a joint project by Juliana Jandre and Vander Viana. Juliana holds a PhD in Education from the Catholic University of Rio de Janeiro (Brazil) and is a teacher of English as a foreign language. Vander holds a PhD in English Language and Linguistics from Queen's University Belfast (UK), and is a lecturer in TESOL and Applied Linguistics. Both contributors have approached the field of English as a foreign language (EFL) from both a theoretical and an applied perspective. Juliana has worked as an EFL teacher for 14 years and has supervised pre-service teachers in their placements since 2011. Vander worked as an EFL teacher for 10 years and, since 2009, has worked with teacher education at all university levels.

Both of us have worked at the Brazilian school where this test was developed and used. The school is a public high school. High schooling in Brazil lasts for three years, and students are generally between 15-17 years old. The test was designed for EFL students in their third and final year at high school. The test was designed for a specific workshop on "Global Englishes," which was attended by 14 students aged 17-18. The students were keen on the topics discussed, and actively participated in class discussions.

The test deals with the theme of appreciating cultural diversity, which is thoroughly discussed in classes. This theme helped to fulfil the school's mission of developing students as fully-fledged Brazilian citizens. In order to be able to live in our globalized society, it is of utmost importance that

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

students learn to avoid cultural and ethnic biases, and to develop their awareness of English as an international language.

More specifically, the test evaluates (primarily) students' reading skills. Most of the questions require students to read both ordinary and multimodal texts to identify what they are about and to relate the texts' content to other spheres of life. Vocabulary is the most important language system being tested, as students need to understand the texts to answer most of the questions. We see vocabulary knowledge as key to understanding texts. There is even an explicit vocabulary question (i.e., Subtest 1, Question d) which asks students to explain two words in one text based on the context in which the words appear.

Two languages are used in the test: English and Portuguese. English is the L2, and Portuguese is the L1. All the input texts and most of the questions are in English. However, students are asked to answer some questions in Portuguese. The rationale here is to check their understanding of the texts since they have to reprocess the information presented in the texts into Portuguese, rather than just copying the English sentences from the input texts.

The Test

An English as a Foreign Language Test for Reading, Writing, and Cultural Diversity Awareness for High School Students

1) Read **Text I** and answer four questions about it.

Text I

HUMAN is a collection of stories and images of our world, offering an immersion to the **core** of what it means to be human. Through these stories full of love and happiness, as well as **hatred** and violence, HUMAN brings us face to face with the Other, making us reflect on our lives. From stories of everyday experiences to accounts of the most unbelievable lives, these poignant encounters share a rare sincerity and underline who we are – our darker side, but also what is most noble in us, and what is universal. Our Earth is shown at its most sublime through never-before-seen aerial images accompanied by **soaring** music, resulting in an ode to the beauty of the world, providing a moment to draw breath and for introspection.

HUMAN is a politically **engaged** work which allows us to embrace the human condition and to reflect on the meaning of our existence.

Source: http://www.human-themovie.org/#about-the-film

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

*Answer 1-a and 1-b in **English**.*

1-a) What is the main objective of Text I?

1-b) What is the main objective of the movie HUMAN?

*Answer 1-c and 1-d in **Portuguese**.*

1-c) "HUMAN brings us face to face with the Other, making us reflect on our lives". Have the excerpts of HUMAN that you watched made you reflect on your own life? Justify your answer.

1-d) We have learned during this year that we do not have to know the exact meaning of words in English to understand what they mean. Choose two of the words highlighted in the text to explain their meaning in their specific contexts.

2) Read **Text II** and answer the following three questions.

Text II

Ethnocentrism is a term applied to the cultural or ethnic bias—whether conscious or unconscious—in which an individual views the world from the perspective of his or her own group, establishing the in-group as archetypal and rating all other groups with reference to this ideal. This form of tunnel vision often results in: (1) an inability to adequately understand cultures that are different from one's own and (2) value judgments that preference the in-group and assert its inherent superiority, thus linking the concept of ethnocentrism to multiple forms of chauvinism and prejudice, including nationalism, tribalism, racism, and even sexism and disability discrimination.

Source: http://www.oxfordbibliographies.com/view/document/obo-9780199766567/obo-9780199766567-0045.xml

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

*Answer 2-a and 2-b in **Portuguese**.*

2-a) How is ethnocentrism defined?

_____

_____

2-b) What is one possible consequence of ethnocentrism?

_____

_____

*Answer 2-c in **English**.*

2-c) Give one example of an ethnocentric way to talk about another culture.

_____

_____

3) Taking into account our class discussions on the so-called "native" and "non-native" speakers of English, answer the following questions in **English**.

3-a) Explain why the English language does not have an "owner" in the current global scenario.

_____

_____

_____

3-b) Brazilians have been using some English words in their daily lives in different ways and contexts. Give at least two examples of such uses and explain them.

_____

_____

4) In 2013, a shampoo brand launched an advertising campaign in Brazil making fun of the way Joel Santana, a soccer coach, speaks English.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Source: https://uolesporte.blogosfera.uol.com.br/2013/07/16/joel-santana-e-seu-ingles-brilham-em-comercial-de-xampu

Answer 4-a in ***English.***

4-a) What do you think about the campaign making fun of the way that Joel Santana speaks English? Explain your answer.

_____
_____
_____

Answer 4-b in ***Portuguese.***

4-b) Explique a influência da língua portuguesa no inglês falado por Joel Santana. Em outras palavras, explique o possível motivo linguístico de um falante brasileiro produzir "donti" e "révi" em inglês. Considere as características formais das línguas inglesa e portuguesa.

_____
_____
_____
_____
_____

English Translation and Answer Key of the Test

For Subtest 4, the caption for the advertising graphic reads:

- *Em:* Portuguese word meaning *In:*.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

- *Donti révi caspa* The first two words make fun of the way that Brazilians would pronounce the words *don't* and *have* in English; *Caspa* is the Portuguese for *dandruff*.

The title of the campaign is "Don't have dandruff", meaning "Get rid of dandruff."

For Subtest 4, Question b, the translation is:

Explain the influence of the Portuguese language in the way Joel Santana speaks English. In other words, explain the likely linguistic reason for a Brazilian speaker to say "donti" and "révi" in English. Consider the formal features of both English and Portuguese languages."

1) Read Text I and answer four questions about it.

Answer 1-a and 1-b in English.

1-a) What is the main objective of Text I?

The main objective of Text I is to describe the movie HUMAN. (Other similar answers were accepted as correct – e.g. to provide the reader with information about the movie HUMAN.)

1-b) What is the main objective of the movie HUMAN?

The movie aims at making its spectators reflect on their lives. (Other reflection foci were accepted – e.g. cultural identities, and the meaning of our existence.)

*Answer 1-c and 1-d in **Portuguese**.*

1-c) "HUMAN brings us face to face with the Other, making us reflect on our lives". Have the excerpts of HUMAN that you watched made you reflect on your own life? Justify your answer.

This question required personal answers from students, and there were no correct or wrong answers. The marking was done on the basis of students' answers being clear, coherent and well-justified.

1-d) We have learned during this year that we do not have to know the exact meaning of words in English to understand what they mean. Choose two of

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

the words highlighted in the text to explain their meaning in their specific contexts.

Text I contained four underlined and bolded words, and students only had to choose two of them. The answers would need to be given in Portuguese, but an English explanation is provided in parentheses.

- core: *parte principal* (the basic or the main part)
- hatred: *ódio* (intense dislike)
- soaring: *extremamente alta* (impressively high)
- engaged: *engajado* (committed – e.g. to a cause)
-

2) Read **Text II** and answer the following three questions.

Answer 2-a and 2-b in ***Portuguese.***

2-a) How is ethnocentrism defined?

Etnocentrismo é definido como o preconceito cultural ou étnico, consciente ou não, em que o indivíduo vê o mundo somente da perspectiva de seu grupo cultural. Esse grupo é considerado como arquetípico e é utilizado como referência para classficação de todos os outros grupos. (The answer is a free translation of the following excerpt: "Ethnocentrism is a term applied to the cultural or ethnic bias—whether conscious or unconscious—in which an individual views the world from the perspective of his or her own group, establishing the in-group as archetypal and rating all other groups with reference to this ideal)."

2-b) What is one possible consequence of ethnocentrism?

(1) Incapacidade de compreender outras culturas diferentes da sua própria de forma adequada
(2) Juízos de valor que privilegiam a cultura própria e afirmam a sua superioridade inerente, relacionando o conceito de etnocentrismo a múltiplas formas de chauvinismo e preconceito como, por exemplo, nacionalismo, tribalismo, racismo e até mesmo sexismo e discriminação por deficiência.
(Students would only need to mention either (1) or (2). The above answers in Portuguese are a free translation of the following excerpt: "(1) an inability to adequately understand cultures that are different from one's own and (2) value judgments that preference the in-group and assert its inherent superiority, thus linking the concept of ethnocentrism to multiple forms of

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

chauvinism and prejudice, including nationalism, tribalism, racism, and even sexism and disability discrimination.")

*Answer 2-c in **English**.*

2-c) Give one example of an ethnocentric way to talk about another culture.

The answer to this question cannot be found in the text. Instead, students should come up with an example of their own. Any cultural stereotype would be a fitting answer – e.g. the British drive on the wrong side of the road; Brazilian partners are clingy; Italians speak loud.

3) Taking into account our class discussions on the so-called "native" and "non-native" speakers of English, answer the following questions in **English**.

3-a) Explain why the English language does not have an "owner" in the current global scenario.
Students were expected to refer to the topics discussed in class such as the fact that non-native speakers of English have outnumbered native speakers, English is widespread around the world, English is used in many different ways and for different purposes.

3-b) Brazilians have been using some English words in their daily lives in different ways and contexts. Give at least two examples of such uses and explain them.
Several answers could be provided for this question. Some examples include:

- "selfie": a picture you take of you or of you and a group of people;
- "shopping": this word is used in Brazil to refer to a shopping mall – not the act of buying;
- "stalkear": the Brazilian Portuguese version for the English verb 'to stalk'.

4) In 2013, a shampoo brand launched an advertising campaign in Brazil making fun of the way Joel Santana, a soccer coach, speaks English.

*Answer 4-a in **English**.*

4-a) What do you think about the campaign making fun of the way that Joel Santana speaks English? Explain your answer.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Students are free to position themselves in any way they wish – e.g. they may find the campaign funny, incomprehensible or offensive. What matters is that they must provide a clear and coherent explanation for their answers.

Answer 4-b in ***Portuguese.***

4-b) Explique a influência da língua portuguesa no inglês falado por Joel Santana. Em outras palavras, explique o possível motivo linguístico de um falante brasileiro produzir "donti" e "révi" em inglês. Considere as características formais das línguas inglesa e portuguesa. (Explain the influence of the Portuguese language in the way Joel Santana speaks English. In other words, explain the likely linguistic reason for a Brazilian speaker to say "donti" and "révi" in English. Consider the formal features of both English and Portuguese languages.")

There are a number of segmental and suprasegmental features that students may discuss in their answers to this question in Portuguese. For example, they may comment on the inclusion of an extra vowel sound at the end of words like "don't" and "have", which end with a consonant sound. This is a feature of Portuguese-speaking users of English since it is uncommon for words in Portuguese to end with a /t/ or /d/ sound.

Contributors' Questionnaire Responses

Section One: Test Planning and Writing

Why did you write the test? What were the purposes of the test?
The primary aims of the test were twofold: To assess and to support student learning. In relation to the former aim, the test helped to identify what the students had learnt throughout the academic trimester. The primary focus was on reading skills since most students were going to undertake a university entrance exam, which basically tests their reading comprehension through multiple-choice questions. The test also focused on students' writing skills as they had to provide written answers to some argumentative questions and to relate the topic of cultural diversity to their lives.

At the same time, the test also was instrumental in stimulating students to consider cultural diversity further. The test was administered at the end of the trimester in which students had engaged in a series of activities (text reading, video sessions, discussions) on a variety of culture-related matters such as the use of English as an international language, the agency of all speakers of English (not only native speakers of English), and ethnocentrism.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

With regard to the first test aim, students were awarded grades based on their answers, thus assessing their learning quantitatively and meeting the formal evaluation requirements of the local high school system. The teaching team also provided students with feedback so as to help them understand how they can improve their learning skills. In this sense, it can be stated that students were provided with comments which indicated why they were awarded a specific grade (feedback) but also with suggestions on how they can improve their performance (feedforward).

How did you decide how many test items to write in total?

The first stage in the development of the present test was the selection of texts to be included in it. This selection was guided by criteria such as topic relevance (the texts needed to relate to cultural diversity), language appropriateness (the texts needed to be understood by third-year high school students at the school where the test was administered) and textual variety (the genres needed to be different and there should be at least one multimodal text).

The main concern was to ensure that students were able to complete the proposed assessment within the available time frame. Because the test was going to be administered as part of the school's test week, they would have a maximum of 100 minutes to complete it. It was therefore important to consider both the time needed to read the texts and items as well as to answer the questions.

A total of ten items was believed to be enough for the time available. This total would also make the scoring easier as each item would be worth one point. After the ten items were created, the test was shared with the two pre-service student teachers who were taking their placement in this specific workshop to do the test and provide feedback on it. Based on their comments, one of the items was split into two (items 2-a and 2-b) since they actually asked two different matters – namely, a definition of ethnocentrism and a possible consequence of it. Presenting these items separately was thought to be beneficial to students: they would not run the risk of forgetting to answer them. In terms of scoring, it was decided that items 2-a and 2-b would be worth 0.5 point each, and the other test items would continue to be worth 1.0 point each. This put the test at eleven items.

Were you concerned at how long the test would take to administer?

This was a primary concern in the design of this test. The amount of time allotted to the test (100 minutes) had been set up by the school, and there was no scope to change it.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

At this school, it is important to consider how much time is needed to complete a test since a test which takes less or more time than anticipated can create issues. For example, if students are able to complete a test in a short amount of time, they will have considerable free time outside the classroom and will need to be supervised. If a test takes longer than anticipated, it may result in students' inability to provide enough evidence of their learning. This would lead to their being awarded a lower grade than the one they would deserve.

Were you concerned at how long the test would take to score?
The amount of time needed to score the test was an important factor to consider. As the school where this test was administered holds a test week, all students sit their tests around the same time. A regular EFL teacher in this context generally teaches five classes with 15 students each, meaning that she or he has to score 75 tests at once.

Given these circumstances, it is ideal to design a test that is easy to score both in relation to the number of questions and to the number of points assigned to each question. As regards the former, a total of ten items are generally included per test (the present one had eleven items because one item was split into two based on feedback from two pre-service student teachers). With regard to the latter, items are generally worth 1 or 0.5 point depending on their complexity. This system saves time when it comes to scoring students' tests.

From students' perspectives, it is also important to ensure that the tests are scored and returned promptly. The sooner tests are returned, the better it is for students as they can more easily understand why they made certain mistakes. The school does not have a formal turnaround policy, but EFL teachers have agreed to provide test feedback within two weeks.

Did you consider having your students take your test on a computer?
Having students take the test on a computer was not considered for two main reasons: technical and security. First, the school has only one small computer lab with approximately ten computers. This lab would not cater for all the 14 students in this class. Because the test was administered in the school's test week, it would also disrupt the entire school planning if EFL students had to leave the test room to take their tests somewhere else while students taking French as a foreign language would stay in the regular classroom.

Second, the use of computers would create problems with test security. The computer lab in the school is organized in such a way that one computer is too close to the other, so one could not be completely sure that students

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

were not cheating. Because the computers are connected to the Internet, students would be able to access any resource they wished online to check their answers, thus compromising the test validity in assessing student learning.

Did you consider making the test an open book test?
This possibility was not considered because one of the main aims of the test was to encourage students to reflect further on the topic of cultural diversity, which had been discussed in class. In this sense, an open-book test would not have been helpful to students as they would not be able to find any answers in published books. The test already contained three texts (two primarily linguistic and one multimodal), providing students with enough material to consult and rely on in order to answer the proposed items.

Given the time limit for test completion (100 minutes), which was decided by the school, it was important to ensure that students' attention was directed to what really mattered. Having an open-book test would probably result in students' being unable to complete the test. They would spend unnecessary time consulting books which would be of little help in their answering the test items.

Did you consider allowing learners to use additional sources such as dictionaries, or their notes, while taking the test?
Students could have been allowed to consult dictionaries if they wanted to look up words that they did not know. However, this would distract them from the main focus of the test. It would also go against the type of reading skill being developed (see, for instance, item 1-d where the meaning of words had to be explained on the basis of the context where they appeared in the text). We also think using dictionaries during the test would take learners' focus from what really mattered, which were the main topics being discussed in the three texts included in the test. The test focused on inference as one of the main reading skills to be mastered by students. They were encouraged not to find out the meaning of each and every word to be able to understand the main points in the input texts.

The possibility of students' consultation of their notes was not considered because this would be of little use to students. The test does not contain questions which encourage students to memorize facts. Instead, it encourages them to interpret, understand, react, and respond to texts.

Did you plan to allow learners to re-take a test for improvement?
Students were not allowed to re-take the test for improvement because it was simply one out of three assessments for the trimester (i.e. the other two

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

assessments were group seminar and class participation). If, for any reason, students did not do well on the test, they could still reach a pass grade overall. For example, a student who scored 3 in the test, 7 in the group seminar and 8 in class participation would have a final grade of 6, which is a pass at this school (the maximum score/grade in each case is 10).

In the Brazilian system, students who do not reach a pass grade have to be offered extra tuition and have to be given the opportunity to take a different test for improvement (Brasil Ministério da Educação, 2016). At this specific school, students who do not reach an overall grade of 6 in a given discipline are invited to attend two extra classes of 50 minutes each to help them work through their difficulties with the teacher and sit another test. The final grade in this case is an average of the original (failing) grade and of the new test score.

The high-school students who participated in the Global Englishes workshop had been encouraged throughout the academic year to re-take their tests informally and autonomously by rewriting their answers. As it was explained to them, this was a way of understanding their mistakes and improving their EFL skills.

What sources did you draw from for your test items?

The test items were designed based on the course objectives (improvement of reading and writing skills, understanding of cultural diversity, recognition of different varieties of English), what learners do in class (watching videos and reading texts on the workshop theme), and their competence (especially their reading competence). For example, items 1-a and 1-b assessed students' reading competence; items 2-a and 2-b focused on cultural diversity; and items 3-a and 3-b dealt with students' understanding of the so-called "native" and "non-native" varieties of English.

The items did not come as a surprise to students as they were used to answering questions that helped them develop their reading skills in EFL and that required them to position themselves in relation to the workshop theme. It is important to ensure that students have already had some practice with the types of questions that they will be asked to answer in a test environment so that they can easily show how much they have learned about the subject matter without struggling with unusual item formats.

Which test item formats do you prefer to use?

The test item format to be used depends on the aim of the evaluation and on the context in which it is applied. The present test was administered to Brazilian high school EFL students undertaking a Global Englishes workshop. The objective of the test was not to check students' mastery of

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

grammatical rules through fill-in-the-blank sentences and/or their lexical knowledge through a cloze exercise, for example. Instead, the test was seen as a pathway for students to evidence their knowledge of EFL, especially reading and writing, through their understanding and engagement with written and multimodal texts.

Some short-answer items were used in the test, such as items 1-a and 1-b where students had to identify the main objectives of Text I and of the movie mentioned in it. A few of the items were more complex and required students to write longer argumentative answers such as item 3-a in which students were asked to explain why the English language has no owner.

What types of learner knowledge do you believe you are capturing in your test? How does that change with test item types you used on the test?
The test helped to assess students' reading comprehension (items 1-b, 2-a, 2-b, 2-c), language/text analysis (items 1-a, 1-d, 3-b, 4-b), and argumentation and/or self-reflection (items 1-c, 3-a, 4-a).

Test items aimed at assessing students' reading comprehension and language/text analysis generally require short answers (e.g., item 1-d focusing on vocabulary inference asks students to choose two words in Text I to explain their meaning in context). Test items on students' argumentation and/or self-reflection demand longer and more complex answers (e.g., in item 4-a, students have to position themselves in relation to an ad in which a famous Brazilian soccer coach's English spoken ability is mocked).

What learner skills do you believe you are capturing in your test? How does that change with test item types you used on the test?
Considering the four traditional skills (listening, speaking, reading and writing), the test focuses more on students' reading and writing skills. The former is tested receptively, that is, students have to understand the three texts included in the test in order to be able to answer the proposed items. Writing skills are assessed productively, which means that students need to be able to express themselves in writing in order to be awarded a pass.

Successful mastery of just one of these skills is not enough. A student who is able to interpret the texts but cannot express themselves in writing will not succeed in the assessment. Similarly, a student who is able to write at length but cannot understand the texts will not be able to answer the questions being asked.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

How did you get ideas on how to score learners' performances (the scoring criteria)?

The scoring criteria varied according to the type of test item. The items which focused on reading comprehension or language/text analysis were scored in a straightforward way. For instance, item 2-a asked students to define ethnocentrism in Portuguese, so students should have been able to show their understanding of the text in English where it is stated that "[e]thnocentrism is a term applied to the cultural or ethnic bias—whether conscious or unconscious—in which an individual views the world from the perspective of his or her own group, establishing the in-group as archetypal and rating all other groups with reference to this ideal" (see Text II).

The items which relied on students' argumentation and/or self-reflection were scored on the basis of students' understanding of the importance of cultural diversity and acceptance of cultural differences (e.g., item 1-c required students to explain whether or not the excerpts from the movie "Human" that they had watched in class made them reflect about their own lives). In these items, the content of the answers was more important than language use (e.g., grammar correction). Students would be awarded all the points allotted to a specific question if they were able to show their ideas in a clear way in spite of potential small mistakes. It was only when mistakes interfered with clarity of expression that students would be penalized. If the answers were vague or incomplete, there would also be some point deduction.

If you used points on a scale for scoring learners' performances, how did you decide on how many points on a scale to use?

The scoring was done comparatively, trying to discriminate among students' answers. To this end, the procedure adopted was to go through all the tests, scoring one test item at a time (e.g., scoring item 1-a in all the tests instead of scoring all the items in one test before proceeding to the next test). Students were awarded all the points allotted to a specific question when they were able to answer test items clearly in English or Portuguese (depending on the instructions provided in the test) using their own words. If students provided vague or incomplete answers, they would be awarded just half the point allotted to a specific item.

Were you concerned about whether you could get a colleague to help you score learners' task performances?

The high school where the test was administered is a primary site for pre-service student teacher placement. At the time, there were two pre-service student teachers taking their placement in the specific workshop where the

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

test was administered. This means that these student teachers observed classes, assisted in the development of class activities, and helped to score the test. A calibration meeting was held before the teacher and the student teachers started scoring the tests so as to agree on the criteria, to ensure consistency across markers and to discuss potential issues. Each test was scored by one student teacher and the teacher, who was in charge of ensuring score consistency across the tests.

Did you plan to give the scoring criteria to the learners for future self- or peer-assessment?
The test was administered in the last trimester of the academic year, which means that the students had already sat two other similar tests before then. The tests for this Global Englishes workshop had been similar in terms of their format and scoring, so students were already aware of the type of questions they were going to be asked and how they were going to be scored. In any case, as usual, after the tests had been scored, the answers were discussed in class and any doubts about the scoring criteria were resolved as well.

Did you seek help from a peer to clarify what your test items or tasks were measuring?
When a pre-final version of the test was ready, it was shared with the two pre-service student teachers who were taking their placement in this workshop. They were asked to do the test themselves and to check the clarity of the test items. Based on their feedback, some minor adjustments were made to the test. For example, items 2-a and 2-b were originally presented as a single item in the pre-final version of the test. In order to make it easier for students to focus on each question being asked, it was decided to present these items separately in the final version.

Did you compare your test to the lessons that learners had?
The items developed for this test were very similar to the activities that the students had undertaken in class. This way, students were already familiar with the types of items being asked.

Did you compare your test to the textbook or other materials learners used?
No textbook was adopted for the Global Englishes workshop. Instead, the teacher developed all the teaching materials for this specific group of students with occasional help from pre-service student teachers. The test abode by the same design features of the class materials and was thus comparable to them.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Are you required to use specific tests in your program?
The school where the test was administered provides teachers with considerable freedom with regard to their pedagogic practice. There is no institutional requirement to use (or to avoid) specific tests. Teachers can exercise their autonomy in deciding what suits best their teaching/learning context. In the case of the present workshop, the test was designed to assess students' understanding of cultural diversity in line with the workshop objectives.

Section Two: Test Administering and Scoring

Were you concerned about test security? What did you do to ensure test security?
Test security was indeed a concern because students could easily cheat on the items checking their reading comprehension and language/text analytical skills (these items had more straightforward answers). The items on argumentation and/or self-reflection were more difficult to be plagiarized because it would be easier to identify similar answers. The concern here was that if students cheated on these questions, one of the main objectives of the workshop (to develop students' understanding of cultural diversity) would not be met as they would not provide their own original answers to the questions being asked.

In order to ensure security, students were asked to skip a row from one another when choosing their seats. This would maintain physical distance between students. The test was invigilated by a school teacher, and students were asked to leave the room once they had finished the test and hand it in to the teacher. They were not given a chance to return to the room, to see their tests again and/or to change any of their answers. The teacher-invigilator only left the room after the last student had completed the test and took all the tests with her.

Were your tests photocopied, or did learners see your test items or test tasks another way, such as on the blackboard?
The test was photocopied at the school one week before its administration. Each student in the class was handed one copy of the test on the day that they sat the test.

How did you deal with learners who missed the test, or who were late for the test?
In these two instances, students have to follow the school regulations. If students miss a test because they did not attend school on that day or arrived

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

later than 15 minutes for class, they were forbidden to enter the test venue. Furthermore, they have to fill in a form at the administrative office within two working days after the test. In this form, they must justify their absence and require a make-up test (generally a new test that teachers create for these specific students).

The forms are analyzed by the pedagogical directors at the school. If the decision is favorable to the students, the pedagogical directors set a day for the make-up test. This day generally falls at the end of the trimester so that students have the opportunity to do a test to substitute any assessment that they may have missed.

How did you prepare learners to take the test? Were there any test items or test procedures that were new to them?

The present test was administered in the last trimester of the academic year. At this point, students had already taken other two similar tests and were familiarized with the test items. As usual, in the class before the test, students were reminded that the test would focus on the topic that had been discussed throughout the semester (cultural diversity) and what was expected of them (to answer reading comprehension questions based on texts and to use these texts as a springboard to position themselves in relation to the topic being discussed). Despite the fact that the test items or procedures were not new to the students, the teacher was present on the day when the test was administered to ensure that the students understood all the items and to answer any doubts that they had.

Did you pilot your test?

The test was not piloted with the high school students because this would not be possible at the school where the test was administered. However, the two pre-service student teachers who were undertaking their placement in this workshop were asked to complete the test and give feedback on it. Although these student teachers cannot be compared to the target group (i.e., undergraduate versus high school students), this could be seen as a form of piloting the test.

Did you write any of the test in the learners' first language? Why?

One of the test items (item 4-b) was written in Portuguese, the students' first language, because it asked students to undertake a cross-linguistic analysis of an ad that made fun of a famous Brazilian soccer coach's spoken ability. The ad had examples of interference (i.e., the way this coach pronounces *don't* and *have*) and of idiosyncratic borrowings (i.e., the use of *caspa* in Portuguese instead of its English counterpart *dandruff*). The question was

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

written in Portuguese and the answer was also to be given in this language due to the complexity of this question to high school EFL students, and due to their need to refer to formal characteristics of both Portuguese and English in their answers. The use of the students' first language in this case was deemed to make the test item more accessible to students.

For classroom tests: How did you accomplish scoring learners' tests?
Each test was scored by two people: a pre-service student teacher and the teacher. Before they started scoring the tests, a calibration meeting was held in which they discussed and decided on the evaluation criteria to be adopted. On this occasion, a test key was prepared for those items which required specific answers (e.g., definition and one possible consequence of ethnocentrism in items 2-a and 2-b). No key was written for items which required less straightforward answers (e.g., item 3-b on two examples of how English words are used in the Brazilian daily context). The scoring team, however, discussed what was expected of students in these freer answers.

Students' names were not hidden as this is not common practice at the school. As a matter of fact, this is not common practice in Brazil for most low-stakes school tests such as this one.

The tests are naturally randomized in this context: they are not kept in any particular order. Because the tests were split among scorers, this also meant that the tests were even further randomly shuffled.

Each of the two pre-service student teachers scored only seven tests, thus having a partial view of student attainment. The teacher, however, second scored all of the 14 tests and had an overall perspective on student achievement. This second scoring procedure also ensured consistency and accuracy across the team.

Going back and changing marks were necessary as the pre-service student teachers were sometimes too strict in their scoring. When correcting Item 4-a, for instance, they considered it wrong when students liked the advertising campaign which mocked the Brazilian soccer coach's English spoken ability, disregarding their argumentation. We had agreed that students should not be penalized for the content of their answers if they were able to provide a rational argument, so some of the tests had to be re-marked.

Do you think your test was reliable? What did you do to check?
Reliability was ensured by sharing the pre-final version of test with two pre-service student teachers who were undertaking their placement in this workshop. These student teachers were asked to take the test themselves,

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

and to provide feedback on the items. Based on their feedback, some minor changes were made to the test for the sake of clarity.

All the students undertook the test under the same external conditions. The test was administered as part of the school's test week. This meant that the students took the test in the same classroom, had the same amount of time (100 minutes), and followed the same procedures (e.g., being invigilated by a school teacher, keeping physical distance between students by having them sit in every other row). The EFL teacher visited the classroom to solve any doubts that students had about the test.

The test results were not submitted to statistical analysis due to the low number of students. With a total of 14 students, it would not be possible to subject the results to robust statistical testing.

Reporting scores

Did you report the scores to learners?
Not only were the scores reported to students, but the tests were also returned to them. One-third of a class meeting was set aside to discuss what the main issues were and how answers could have been improved. Students were asked to check that their scores were accurate, that is, to confirm that the individual item scores added up to their overall test scores. Some students had follow-up questions about their tests, which were dealt with individually on a one-to-one basis.

What was your goal in reporting the test scores to learners?
The test fulfilled both a formative and a summative role. It was formative in the sense that some of the items fostered, for instance, students' self-reflection. The feedback provided for all of the test items was also formative. It helped students identify how they could improve their reading and writing skills. At the same time, the test had a summative aspect. It evaluated students' learning attainment in quantitative terms. This summative aspect helps to show students and other key stakeholders (e.g. parents, the school principal) how much progress students are making in EFL.

Did you teach learners how to interpret their test scores?
The test was administered to high school students in their final year, which means that most of them have been at this school for at least three years (the entire duration of high schooling in Brazil). The students were already used to the interpretation of test scores. The pass score at the school is 6. A score

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

below 6 indicates that students' performance needs improvement, and a score above it shows that they are doing well.

Did you report the scores to anyone else?
The final EFL grade (the average of the three assignments for the workshop) was printed in the school report alongside the grades students were awarded for the other 12 school subjects. This report is given to students, who must show it to their parents. If a student does not reach a passing grade, parents are invited for a meeting with the pedagogical director and coordinator of the relevant subject to discuss the situation and to find ways to support the student.

Did you spend time explaining scores, or answering learners' questions about scores in or out of class?
One-third of the class when the tests were returned to students was set aside to explain the scoring rationale. Students had a chance to clarify their doubts in front of the group if they so wished, and a few decided to talk to the teacher on a one-to-one basis.

How quickly did you report scores to learners? Was speed a priority?
The high school where the test was administered does not have a policy on the maximum amount of time that teachers can take in order to report scores to students. However, students usually feel anxious about their test scores. Taking this into account, the EFL teachers at the school have agreed on a two-week turnaround policy for tests. Pedagogically, it makes sense to return tests as soon as possible so that students remember the difficulties they had faced and are able to act on them. On this specific occasion, the scores were reported to students on the class immediately after the one in which they had sat the test. It was therefore possible to discuss the test extensively and use it in a formative way, that is, as a tool to enhance student learning.

Section Three: Using Test Scores

Cut scores

How did you decide which scores were passing or failing scores (cut scores)? How did you decide which scores meant a specific grade on a test?
The high school where the test was administered stipulates a cut overall grade of 6 (out of 10), which is valid for all disciplines. This minimum threshold does not need to be met for each assessment, though. This means

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

that students can compensate a score lower than 6 on the test if they achieve a score higher than 6 on class participation and/or group seminar. They pass the trimester if their overall average is 6. Students should ideally try to reach the cut overall grade every trimester; however, they may still pass the year if they reach an annual average of 6.

How did you use learners' scores from this test?
The test scores were primarily used in the teacher's evaluation of student learning and in her self-evaluation of her pedagogical practice. Had all students failed the test, this would have been a clear indication that something had gone wrong with the classes. Indirectly, the test scores were also used by the school as an indicator of each student's passing or failing EFL in that trimester. This use was indirect because the final scores that were passed on to the school administration were an average of the scores that students were awarded for the three proposed assessments in the trimester (i.e., class participation, group seminar, and test).

The test scores were used by the students as a quantitative indicator of their learning achievement in EFL in that specific trimester. When compared to their previous scores, the test scores show how much/little students have progressed over time.

The test scores may also have been used by parents as a way of finding out how well (or how poorly) their children did in EFL. This use is more difficult to tell as one cannot know whether students choose to tell their parents each and every assessment score or whether they decide to report only on their final trimester grade. It is the latter that the school administration informs parents about it, though.

What was the role of the test score in determining learners' grades?
At the school where the test was administered, teachers can decide on the number of subtests that their students will have to take. The only pedagogical guideline is that students must be assessed at least twice per trimester. This practice ensures that students have a chance to do well overall even if they face an issue with one of the proposed assessments. There is little restriction on the type of assessment to be used. Teachers are free to choose from, for instance, in-class/take-home tests, group presentations or essays. However, the school has a test week, that is, a week in which students come to the school to be assessed on their knowledge of each of the 13 subjects that they study (i.e., Arts, Chemistry, Geography, Geometric Design, History, Foreign Language, Literature, Physical Education, Portuguese, Mathematics, Philosophy, Physics, Sociology). As

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

students cannot be dismissed on this day for a specific subject, there must be an assessment for the students to undertake during this week.

In the trimester when the test was administered, students were also evaluated through two other means: group seminar and class participation. Delivered in pairs or trios, each group seminar lasted for 15 minutes. The groups had to select one English-speaking country (Hong Kong, India, Jamaica, Nigeria and South Africa were some of the chosen countries), research its history and culture, and present the results orally to the class. The participation assessment encouraged students to contribute to classes throughout the trimester, rewarding them for activities like task completion, peer feedback and involvement in class discussions.

The three proposed assessments (the test presented here, the group seminar, and class participation) complemented each other in the following ways: 1.) There was a mix of individual (test and class participation) and group assessment (group seminar); 2.) The assessments required different interactional patterns – individual (test), pair/trio (group seminar), and class (participation); 3.) There was a variety of assessment contexts, from the more relaxed classroom environment (class participation), to the more traditional exam setting (test); 4.) Students were evaluated on a single occasion (test and group seminar) and longitudinally (class participation); 5.) The proposed assessments took place on different days so as to avoid potential personal circumstances from affecting students' performance in the module; and 6.) All three assessments had both a formative and summative role. They provided students with guidance on how they could improve their performance while also measuring how they had done up to that stage.

These three assessments provided students with opportunities to show that they had met the course objectives – both the linguistic-oriented ones (reading and writing skill improvement, oral argumentation development) and the social-oriented ones (recognition of different varieties of English, appreciation of cultural relativism, understanding of cultural diversity).

How much weight did you give your test? How did you decide? Were other measures used to decide learners' grades, besides your test?

The test included in this chapter was administered in the last academic trimester during which students were also evaluated through a group seminar and class participation. In all three cases, students' scores ranged from zero to ten with decimals being used (e.g., 6.2 or 8.1 were valid scores).

Although the seminar was delivered in pair or trios, each student was evaluated separately on the basis of his/her knowledge of the seminar topic and of his/her presentation skills. The class participation score was jointly

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

decided by the teacher and the two pre-service student teachers, who had observed the students' task completion and contribution to classes throughout the trimester. Because there were only 14 students in class, this close observation was possible.

Students' trimester grades were arrived at by averaging their group seminar, test and participation scores. These three evaluations had the same weight in the final grade (i.e. the three individual scores were added up and divided by three) as they measured different aspects of students' development.

Both the group seminar and the participation assessed students' speaking and listening skills. The group seminar provided students with a chance to plan what they were going to say while the participation assessment encouraged them to contribute in a more spontaneous way to the classes. Another difference between these two spoken evaluations is that the group seminar was punctual (the students were evaluated on a single occasion) while the participation was inherently durative (the students were evaluated throughout the trimester in a longitudinal way).

The test presented here differed from the other two assessments in that it focused on students' reading and writing skills. The questions required students to read (both linguistic and multimodal) texts to identify general and specific information and to relate the contents of the selected texts to other spheres of life. Vocabulary was the most important language system being tested as students needed to understand the gist of each text in order to answer most of the questions.

Reporting scores

How did you report scores to learners? Was timeliness of concern to you?
Students learned about their scores in the class immediately after the one in which they had sat the test. Their respective tests were returned, and the test items were discussed in class. The discussion centered on what the most common issues were and how the answers could have been improved. The timeliness of feedback was a key factor in ensuring a successful discussion. Students could still remember the potential difficulties that they had faced and could actively use the feedback provided in a formative way.

Did you hand the test back to learners? Did the learners get to keep the tests?
The tests were handed back to the students, who could keep them.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you offer feedback to individual learners in addition to their test scores?
Students received both written and spoken feedback. Not only were the tests annotated with remarks on what they had done well or could have done better, but time was also set aside in class to discuss the test items. Some students sought additional clarification, and this was provided on a one-to-one basis.

Did you teach learners to interpret their scores?
Because the test was administered to third-year high school students, they were already familiarized with the scoring system adopted at the school. Most of the students had already studied at this school for at least three years. They were already aware that the passing grade is a score of 6. A score below 6 shows that they need to improve their performance, and a score above it indicates that they are doing well.

Using test scores

Do you feel you learned what you needed to learn from the test results about what learners could and could not do, or what they knew or did not know?
The test was successful in evidencing students' reading and writing skills. They had to unpack the information provided in the three input texts, process this information and repackage it in order to answer the test items. The test also showed students' ability to relate the content of the selected texts to other spheres of their lives, and it indicated how accepting of cultural diversity students are.

Did your test change how learners studied?
The test potentially provided students with another perspective on studying. They did not need to memorize facts or concepts in order to reach a passing grade. Instead, they needed to show functional engagement with EFL, that is, they needed to use it as a means to understand the points made in the selected texts and to convey their own thoughts through writing.

We wonder, though, whether the test caused washback as it was only developed after most of the classes had been taught. The classes were designed in such a way to meet the aims of the proposed workshop, and the test was coherent with the activities undertaken in class (as well as the workshop aims). In line with Christison & Murray's (2014, pp. 60–61) suggestion that "assessment must be aligned to the goals and objectives of the curriculum", the test seems to have been more influenced by the teaching than vice-versa.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

As feedback was provided in the class immediately after the one in which students had taken the test, students were able to relate to the discussion held in class and learned how they could improve their reading and writing skills in EFL.

Did you spend time going over the test in class?
The test was discussed in the class immediately after the one in which the students had taken the test. Each item was discussed, and feedback was provided to the entire class by identifying trouble points and commenting on how answers could have been improved. Students participated in the discussion by offering answers and asking about the mistakes that they had made.

Did learners ask you about the test itself (not the test scores) outside of class? If so, what did they want to talk to you about?
Some students sought further clarification about the test scores during the class on an individual basis. While they had been able to follow the class discussion, they wanted to understand why some of their specific answers were not correct. They decided to do so on a one-to-one basis because they did not want to read their answers out loud in class.

Did learners' test scores change your teaching?
In this specific case, the scores did not bring about a change in pedagogical practice because the test was the last evaluation in the last trimester for the last year in high school. However, some lessons were learned with this assessment (e.g., a few students still held some bias against English language varieties other than standard American and British ones). This will be considered the next time that this workshop is offered.

If you could turn back time, what would you change about your test? What would you change about your test administration?
Students thought that the test was fair because it dealt with topics that had been discussed in class, the items were familiar to them and the skills required of them had been developed throughout the academic year. They did not suggest any changes to the test.

If the test were to be used again in the future, it would be worthwhile making it more challenging to the students. They finished the test ahead of the time (they originally had 100 minutes to complete the test, an amount of time decided by the school administration), and the scores were quite good overall. The test could have better discriminated among students' reading/writing skills.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

In terms of specific changes, the texts could be more complex. Either the entire original texts could be used or other more appropriate texts could be chosen. It would also be fitting to indicate to students how many points each item is worth. This way, students can decide the order in which they want to answer the test and which items they wish to spend more time on.

Section Four: Evaluating and Reviewing your Answers

To what extent do you think you've described recurrent patterns in your work with tests?
The above answers describe a recurrent pedagogical practice when it comes to high school EFL teaching with a focus on reading/writing skills. The practice would only be different if the focus of the course were different (e.g., a course which focuses on the development of students' speaking skills or grammatical knowledge).

To what extent is your test here an innovation, or something new, for you?
This test is not an innovation for us because we are used to designing/implementing courses which focus on reading and writing skills for Brazilian high school EFL students. However, it is an innovative test in the context where it was administered as high school teaching in Brazil would generally focus on grammar.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

A Final Exam on Contextual English Grammar for Pre-service Teachers of English

Ferit Kılıçkaya
Burdur Mehmet Akif Ersoy University

Introduction

I am a teacher trainer currently working in the department of foreign language education at Burdur Mehmet Akif Ersoy University in Turkey, where students are trained to be teachers of English as a Foreign Language. During their four-year undergraduate education, students are provided with a foundation in English language and literature, teaching language skills and methodology, linguistics and educational sciences. They can be teachers of English in primary and secondary institutions to teach primary, secondary, and high school students. They can also work as lecturers of English at the Schools of Foreign Languages provided that they earn an M.A. degree in their field of study.

The attached test is the final exam for the class, Contextual Grammar I, a required course for freshmen. This course aims to refresh students' linguistic competence in English, to review the basic and advanced linguistic structures such as noun clauses and reduction of adverb clauses frequently used in texts students are expected to encounter, to create an awareness of the relationship between the linguistic structures and lexical items and the meanings, and to analyze the language structures within the framework of a context of meaning.

Upon successful completion of the course, students ought to recognize context and use of structure in foreign language teaching, to explain the relationship between context and grammar, to acquire a good understanding of a variety of structures with correct grammar, leading to a comprehensive knowledge about English grammar and to be able to put this knowledge into use in academic contexts as well as other situations. Students are expected

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

to attend each class regularly, and absences cannot exceed 20% of class hours.

The class for Contextual Grammar I includes 62 English as a Foreign Language (EFL) freshmen (45 female and 17 male). The students are all native speakers of Turkish with an average age of 19.3, ranging from 18 to 21. All of the students, who are pre-service teachers, are graduates of high schools and have taken the required university entrance examination with a section on language assessment that included multiple-choice questions on grammar, reading, and vocabulary. The main coursebook used is *Grammar dimensions 4: Form, meaning, and use,* 4th ed. (Frodesen & Eyring, 2007). Moreover, students are advised to consult various websites, such as http://grammar.ccc.commnet.edu/grammar/textonly.htm and http://www.chompchomp.com/menu.htm, and to review what they have learned in class and/or advance their current knowledge. The grading is based on two exams: a midterm (40% of their grade) and a final exam (60% of their grade), which include questions similar to the ones provided on the workbook/student book. The midterm and final exams will test students' knowledge of the material covered in class and consist of multiple-choice questions based on the exercises on the book.

The attached test is the final exam paper for the Contextual English Grammar, which was used at the end of the spring semester. In order to increase the content validity of the test, 66 questions are included in the test, and the maximum score to be obtained from the test is 100. The questions are prepared using alternate test item formats (fill in the blank, short answer, rewrite, etc.) based on what learner knowledge and skills the teacher wishes to capture, and the content and item types found in the coursebook. Since the test is a final test, it includes a variety of topics in English grammar, including 1) relative clauses, 2) reduction of relative clauses, 3) passive voice, 4) complex passives, 5) verb tenses, 6) subject-verb agreement, and 7) ambiguous references.

The Test

BURDUR MEHMET AKİF ERSOY UNIVERSITY
DEPARTMENT OF FOREIGN LANGUAGE EDUCATION
CONTEXTUAL GRAMMAR I FINAL EXAM
Time allowed: 70 minutes

Student's Full Name:	**Student's ID Number:**	**Signature:**

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Instructions: Read the instructions carefully, write legibly, and express you clearly. Points will be deducted for incorrect spelling or grammar, and unacceptable expressions.

(A) Please fill in the blanks of the following two passages using the relative pronouns, *which*, *who*, *where*, *whose*, and *whom*. You may use some of the relative pronouns more than once, while you may not need some (Each is 1 pt.).

(A) Last Sunday, David Cameron marched through Paris in solidarity, so it seemed, with those **(1)** _____ stand up for free speech. Anyone **(2)** _____ thought he meant it must be crying out, 'Je suis un right Charlie!' Hardly had the march finished than the Prime Minister had rediscovered his other side: the one **(3)** _____ reacts to terror by threating yet more surveillance, more state control. He has promised to revive, in the Conservative manifesto, the 'snooper's charter' **(4)** _____ would allow the state to retain indefinitely information about very email we ever send, every telephone call we ever make.

From the article, Cameron vs Charlie, published in The Spectator (January 17, 2015).

(B) Children begin to lose gradually the free and easy ways of the primary school, for they sense the need for a more cautious approach in the secondary school, **(5)** _____ there are other pupils. Secondary staff, **(6)** _____ seem to devote most of their time to above average pupils, suffer from the pressures of academic work and seem to have less time to stop and talk.

(C) The effect of a snake bite varies considerably. It depends upon several things, one of **(7)** _____ is the body-weight of the person. The heavier the person is, the less harmful the bite is likely to be, **(8)** _____ is why children suffer far more seriously from snake bites than adults.

(D) The forest **(9)** _____ man takes his timber from is the tallest, most impressive and essential plant community on Earth. Considering man's brief life, it appears permanent and unchanging, except for the seasonal growth and fall of the leaves. No green landscape **(10)** _____ we see today has been forest for all time.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

(B) Please re-write the following sentences putting the focus on the recipient rather than the agent (Each is 1 pt.).

11. Behaviorists can explain personality in terms of stimuli, responses, and reinforcement.

12. Many people have held demonstrations in recent years in protest against the level of pollution.

13. Social factors affect all aspects of man's behavior.

14. Teachers should consider the needs of the learner when specifying relevant goals.

15. The bulb flashed when a fencer made a hit.

(C) Use the correct form of the verbs given in parentheses (Each is 1 pt.).

In dyslexia, less brain tissue not to blame for reading difficulties
Adapted from
http://www.sciencedaily.com/releases/2014/01/140114202913.htm

In people with dyslexia, less gray matter in the brain has been linked to reading disabilities, but now new evidence **(16)** _____ **(suggest)** this is a consequence of poorer reading experiences and not the root cause of the disorder. It **(17)** _____ **(assume)** that the difference in the amount of gray matter might, in part, **(18)** _____ **(explain)** why dyslexic children **(19)** _____ **(have)** difficulties correctly and fluently mapping the sounds in words to their written counterparts during reading. The findings from anatomical brain studies that **(20)** _____ **(conduct)** at Georgetown University Medical Center (GUMC) in the Center for the Study of Learning led by neuroscientist Guinevere Eden, DPhil, **(21)** _____ **(publish)** yesterday in The Journal of Neuroscience. The study **(22)** _____ **(compare)** a group of dyslexic children with two different control groups: an age-matched group which **(23)** _____ **(include)** in most previous studies, and a group of younger children who **(24)** _____ **(match)** at the same reading level as the children with dyslexia. The dyslexic groups **(25)** _____ **(show)** less gray matter compared with a control group matched by age, consistent with previous findings. However, the result **(26)** _____ **(replicate, not)**

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

when a control group matched by reading level was used as the comparison group with the dyslexics. "This suggests that the anatomical differences reported in left hemisphere language processing region **(27) _____ (appear)** to be a consequence of reading experience as opposed to a cause of dyslexia," says Anthony Krafnick, PhD, lead author of the publication. The work also helps to determine the fine line between experience-induced changes in the brain and differences that are the cause of cognitive impairment. For example, it **(28) _____ (know)** from studies in illiterate people who attain reading skills as adults that this type of learning **(29) _____ (induce)** growth of brain matter. Similar learning-induced changes in typical readers may result in discrepancies between them and their dyslexic peers, who have not enjoyed the same reading experiences and thus **(30) _____ (undergo, not)** similar changes in brain structure.

(D) Reduce the relative clauses in the following sentences to change them into adjective phrases if possible. First, underline the word(s) that can be reduced and then write the reduced form. If not possible, write "impossible to reduce" under the sentence (Each is 2 pts.).

31. Middle East Technical University, which was founded in 1956, is an English-medium university.

32. The points which have been raised at the meeting are quite relevant.

33. The new government must address the problems which underlie racial unrest.

34. Dr. Yaka, who was ashamed of his behavior, apologized.

35. The traffic rules that drivers ignore endanger road safety and human life.

(E) Rewrite the following complex passive sentences starting with the word(s) given (Each is 2 pts.).

36. It is stated that fierce fighting is continuing along the southern front.
Fierce fighting

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

37. It is believed that economic stability is crucial for political stability.
Economic stability

38. It was believed that he was against changes whatsoever.
He

39. It was reported that the President had suffered a heart attack.
The President

40. It is said that he was very rich before he went bankrupt.
He

41. It is thought that the woman have been killed out of jealousy.
The woman

(F) Please rewrite the following sentences in order to combine them into one sentence by using an appropriate relative determiner/ pronoun except. Please pay attention to punctuation as well as the difference between defining (restrictive) and non-defining (non-restrictive) clauses (Each is 2 pts.).

42. There are thousands of kinds of bacteria. Many of them are harmless to humans.

43. Shakespeare wrote many plays. The most famous of them is Hamlet.

44. Ferit passed all his exams. This surprised his friends.

45. He is the president of a country of 90 million people. In this country, only 20% of women are in legal employment.

46. The adder is a venomous snake. Its bite may prove fatal to humans.

(G) Read each sentence. Write OK after sentences that are well-formed, parallel, and not repetitious. Rewrite the rest for concise, formal style (Each is 2 pts.).

47. I have neither read the book nor seen the movie.

48. My brother either eats fish or eats chicken.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

49. Not only did they rob him of his money but they also beat him badly.

50. Not only did Imelda get into the honors program, but also Nelson.

51. I know neither where Alma is nor what she is doing.

(H) Please fill in the blanks using appropriate forms of the verbs given in parentheses (Each is 1 pt.)

Customer Service: Hello, what **(52)** _____ (seem) to be the problem, Mr. Kara?

Mr. Kara: I **(53)** _____ (send) in my money for the subscription to your magazine, *WithoutSleep*, several months ago, but so far I **(54)** _____ (receive, not) any issues of the magazine.

Customer Service: I'm really sorry to hear that. Unfortunately, we are experiencing a problem with one of the main computer servers in our company. So, our subscription system **(55)** _____ (function, not) at the moment. However, our computer technicians **(56)** _____ (work) very hard to fix the problem at the present time. We **(57)** _____ (start) your new subscription as soon as the problem is resolved.

Mr. Kara: Thank you so much. Looking forward to it.

(I) Please circle the correct one (Each is 1pt.)

58. Neither they nor I **(are/am)** responsible for the damage caused.
59. The leisure activity of reading books **(seem/seems)** to be thriving, as indicated by polls and other sources.
60. In Great Britain, a study of people's diaries about reading habits **(reveal/reveals)** that most people reported they were reading a book at some point during a three-month period.
61. In other words, a reference book, and not literature or serious nonfiction, **(was/were)** the only kind of book read.
62. A recent survey of literary reading in America by the National Endowment for the Arts concludes that literary reading among almost all groups of Americans **(is/are)** declining.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

J. Ambiguous reference occurs when a pronoun can refer to more than one antecedent. Please explain the problems in the following sentence if there is any and suggest a solution. If you think that the sentences are OK, write "OK" under the sentence (4 pts.)

63. If a baby does not thrive on raw milk, boil it.

64. Ahmet never argues with his father when he's drunk.

K. One of your students asks you to explain the difference in meaning, if there is any, between the following two sentences. How will you respond? (7 pts.)

a. The students who did not study failed.
b. The students, who did not study, failed.

L. Some English language learners, especially Turkish learners of English, might claim that the following two sentences mean the same. Please explain the difference in meaning between two sentences given below (6 pts.).

a. She worked at the company for two years.
b. She has worked at the company for two years.

.:: END OF THE EXAM. PLEASE CHECK YOUR ANSWERS ::.

Answer Key of the Test

(A)

1.) who 2.) who 3.) which 4). which 5.) where 6.) who 7.) which 8.) which 9.) which 10.) which

(B)

11.) Personality can be explained in terms of stimuli, responses, and reinforcement (by behaviorists).
12.) Demonstrations have been held in recent years in protest against the level of pollution (by many people).

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

13.) All aspects of man's behavior are affected by social factors.
14.) The needs of the learner should be considered when specifying relevant goals.
15.) The bulb flashed when a hit was made (by a fencer).

(C)
16.) suggests
17.) is assumed / has been assumed
18.) explain
19.) have
20.) were conducted
21.) were published
22.) compared
23.) included
24.) were matched
25.) showed
26.) was not /wasn't replicated
27.) appear
28.) is known
29.) induces
30.) have not/haven't undergone

(D)
31.) Middle East Technical University, founded in 1956, is an English-medium university.
32.) The points raised at the meeting are quite relevant.
33.) The new government must address the problems underlying racial unrest.
34.) Dr. Yaka, ashamed of his behavior, apologized.
35.) Impossible to reduce

(E)
36.) Fierce fighting is stated to be continuing along the southern front.
37.) Economic stability is believed to be crucial for political stability.
38.) He was believed to be against changes whatsoever.
39.) The President was reported to have suffered a heart attack.
40.) He is said to have been rich before he went bankrupt.
41.) The woman is thought to have been killed out of jealousy.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

(F)
42.) There are thousands of kinds of bacteria, many of which are harmless to humans.
43.) Shakespeare wrote many plays, the most famous of which is Hamlet.
44.) Ferit passed all his exams, which surprised his friends.
45.) He is the president of a country of 90 million people where / in which only 20% of women are in legal employment.
46.) The adder is a venomous snake whose bite may prove fatal to humans.

(G)
47.) OK
48.) My brother either eats fish or chicken.
49.) They not only robbed him of his money but also beat him badly.
50.) Not only Imelda but also Nelson got into the honors program.
51.) OK

(H)
52.) seems
53.) sent
54.) have not /haven't received
55.) is not/isn't functioning
56.) are working
57.) I'll /will start

(I)
58.) am
59.) seems
60.) reveals
61.) was
62.) is

(J)
63.) If a baby does not thrive on raw milk, boil it.
The ambiguity is related to the use of the pronoun "it." "It" might refer to the words "a baby" and "raw milk" at the same time. Therefore, in order to remove the ambiguity, one can rewrite the independent clause as "boil the milk."

64.) Ahmet never argues with his father when he's drunk.
Ambiguous reference in this sentence is related to the use of "he," which refers to more than one antecedent "Ahmet" and "his father." In order to

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

remove the ambiguity here, one can rewrite the dependent clause as "when Ahmet's drunk" or "when his father's drunk."

(K)
a. The students who did not study failed.
b. The students, who did not study, failed.

Sentence (a) means that there are some students and these students did not study and failed. However, the other students passed. Sentence (b) means that all the students did not study and they all failed.

(L)
a. She worked at the company for two years.
b. She has worked at the company for two years.

In sentence (a), we understand that she no longer works at that company. However, in sentence (b), it is indicated that she started working at the company two years ago and she is still there. The focus is on the duration.

Contributor's Questionnaire Responses

Section One: Test Planning and Writing

Why did you write the test? What were the purposes of the test?
I wrote the exam mainly to encourage learners to review the language structures that were practiced in context throughout the semester. The second reason was to award a part of a grade. This exam accounted for 60% of their overall grade in the class. However, although I shared the results and the exam questions together with the correct answers, it was not possible to give each student detailed feedback.

How did you decide how many subtests to write? Was there some correspondence between a subtest and a course objective, or was there some other reason to create the subtests you did?
I determined the subtests based on the language structures covered during the class: 1) 'relative clauses' in Sections A, F, and K, 2.) 'passive voice' in Section B, 3.) 'verb tenses' in Sections C, H, and L, 4.) 'relative clauses' in Section D, 5.) 'complex passives' in Section E , 6.) 'parallel structure' in Section G, 7.) 'subject-verb agreement' in Section I, and 8.) 'ambiguous references' in Section J. I can say that the subtests were mainly based on the content of the coursebook used. There was some correspondence between a

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

subtest and the course content; however, I did not much focus on the course objectives specifically as the objectives were too general to write actual items from.

Did you write as many items or test tasks as you needed, or did you write more than you needed, then pare the number down?
I first determined the possible number of questions that could be asked during the exam duration, which was 50 minutes at most, as there were other exams to be conducted in the exam rooms. Therefore, the exams did not last more than 50 minutes. Then, based on the contents of the coursebook and the corresponding weight of each structure in the book, I divided the number of questions by the number of the structures. As a final step, I wrote the questions based on this number. In this way, I did not write more than I needed.

How did you decide how many items to write for each subtest?
I determined how many items to write for each subtest based on the weight of the structure. By this, I mean the amount of explanation in addition to the examples provided.

Did you have one version of your test, or did you create a second equivalent version?
I had only one version of my test.

Were you concerned at how long the test would take to score?
It was a real concern since there were over 60 students in the class and the questions were not in multiple-choice formats but prepared using alternate test item formats (fill in the blank, short answer, rewrite, etc.), which would make the scoring take longer and a little bit complex considering the different scores that needed to be given for different types of questions. In addition, using such item items might lead learners to come up with alternate answers that I would have to consider for each question.

Were you concerned how you might use the test items themselves for learner feedback?
To some extent, yes. Since it was a final exam and we did not have the opportunity to come together after the exam, I shared the results, the exam questions, and the correct/possible answers via email and other social networking tools. Beyond that, however, it was not possible to give each student detailed feedback.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you consider having your students take your test on a computer?
I could not consider having my students take the test on a computer because there was not any computer lab available for that number of students to take the exam at the same time. Moreover, if there were, I would have several other concerns such as technical issues.

Did you consider making the test an open book test?
I did not consider making the test an open book test as the questions did not require them to think critically and prove their own thoughts based on the concepts, which, for example, can be done in a reading or writing class.

Did you plan to allow learners to re-take a test for improvement? The same test or a different test?
It was not possible for learners to re-take the test for improvement. However, for those who failed the final exam, I prepared an equivalent test as a re-take exam.

What sources did you draw from for your test items?
I used several sources. However, most of the test items were based on the language structures coursebook content and coursebook activities. However, in order not to encourage students to memorize the items on the activities on the coursebook, I did not ask the same items on the final exam. I searched the Internet for academic texts as well as several e-books that I have for the sentences and paragraphs which would exemplify the structures covered. Then, I rewrote or transformed these sentences and paragraphs into questions.

Did you seek help from a peer to clarify what your test items or tasks were measuring?
I could not ask my colleagues to check my items in the test as they had an overloaded teaching schedule. However, I asked two of my M.A. students in our graduate course in ELT (English Language Teaching) to evaluate my items based on the course content. Based on these graduate students' comments and suggestions, I introduced several changes to the test, especially in the instructions.

Are you required to use specific tests in your program?
I am not required to use any specific tests in my program. Actually, everybody is free to prepare any exam they consider appropriate to assess their learners depending on the type of course and content.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you use items or ideas or content from previous tests?
I actually used several items and content from my previous tests that I used for previous classes. As a rule of my own, I change the items in my exams when I teach the course again. Since I share my test and answers with my learners, I do not let my new learners have advantages over others by being asked the same questions.

How did you change the parts you decided to keep?
I used the texts I kept in a different way. For example, if I used a text to set up items on relative questions, this time I used the same to text to set up items on reduction of adverbial clauses, by changing the structures.

What did you add that was new, or different?
Since my learners are also pre-service teachers of English and they will need to deal with teaching English grammar in their future careers, I also included several test tasks in which they have to respond to imaginary situations in which they will explain the difference between similar structures to confused learners. The following exemplifies these situations:

> Some English language leamers, especially Turkish learners of English, might claim that the following two sentences mean the same. Please explain the difference in meaning between two sentences given below:
>
> The students who did not study failed.
> The students, who did not study, failed.

Section Two: Test Administering and Scoring

Were you concerned about test security? What did you do to ensure test security?
I was a little bit concerned about the test security as there was only one proctor for sixty students in a classroom and all of them were asked the same questions in the same order. Preparing different versions (changing the order of the questions, for example) was an option; however, since I believe that that might also affect my students' performance, I did not do that. I was also in the classroom during the exam so that I could help the proctor.

Were your tests photocopied, or did learners see your test items or test tasks another way, such as on the blackboard?
I photocopied the tests myself before the exam date.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

How did you deal with learners who missed the test, or who were late for the test?
There were no students who missed the test. If there were, they would have to write a petition regarding this with a valid excuse to be accepted by the faculty board for a make-up exam. No late students were accepted for the test after 15 minutes. Also, after 15 minutes, no students were allowed to leave the classroom. Moreover, the late comers were not granted any extra time.

How did you prepare learners to take the test? Were there any test items or test procedures that were new to them?
There were not any test items or test procedures that were new to my students. I tried to do my best to write and use the items that my students were already familiar with. Moreover, I did many similar activities with the same items formats in the classroom.

Did you pilot your test?
Actually, I did not pilot my test. However, I obtained my MA students' views on the complexity and difficulty of the items.

Did you write any of the test in the learners' first language?
I did not write any of the test in the learners' first language, as the medium of instruction in the department is English and we try to encourage our students to use English in spoken or written form.

For classroom tests: How did you accomplish scoring learners' tests?
I did not score learners' tests twice for accuracy; however, I carefully checked the students' answers and gave breaks frequently not to allow my tiredness to affect my scoring in any way. I did the scoring myself based on the test key that I prepared and considered alternate answers during scoring by writing down any possible answer that might be accepted during scoring. I did not hide learners' names as I scored, but tried not to look at names. I did not randomize learners' tests for scoring in any way. Sometimes I had to go back and change my marks on previously scored tests in response to problems I found while scoring tests later in the process.

Did you put learners' responses to items into a spreadsheet for further analysis? Did that process help you catch scoring accuracy problems?
I did not put learners' responses to items in any way for further analysis. However, I wrote down the problematic answers or the questions that seem to cause problems as many students could not answer correctly.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Do you think your test was reliable? What did you do to check?
I think my test was reliable. I could not ask another lecturer or teacher to go over my test before administered it. However, I asked several of my MA students to check the items. Based on their responses, I introduced several changes such as simplifying the language and the selection of the examples. All learners took the test in their classes where they had their regular meetings with their lecturers. The conditions, therefore, were the same for all students. I used item analyses using item facility.

Reporting scores

Did you report the scores to learners?
I reported the scores to learners online on the student affairs information system of the university. The students easily interpreted their test scores as this system provided the mean and other basic statistics. I did not report the scores to anyone else. Moreover, I did not spend time explaining scores as it was a final exam and there was no class after the final exams. However, as I told my students to get in touch with me through e-mails or any other possible way, I answered learners' questions out of class, mostly through emails. I reported scores to learners in seven days after the exam. Timeliness is real concern as I think that learners should be provided with the results as soon as possible.

Section Three: Using Test Scores

Cut scores

How did you decide which scores were passing or failing scores (cut scores)? How did you decide which scores meant a specific grade on a test? Did your institution stipulate cut scores?
I did not decide cut scores as my institution stipulated cut scores automatically when I was finished with entering learners' scores on the student affairs information system.

How did you use learners' scores from this test?
These scores were used to award a part of their final grade.

Did anyone else use the test scores?
No one else used the test scores.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did learners' test scores have any positive or negative consequences for you, in terms of your institution?
These test scores did not have any positive or negative consequences for me.

What was the role of the test score in determining learners' grades? Consider: How much weight did you give your test? How did you decide?
The results of this test accounted for 60% of the learners' overall grade in the course. Therefore, the results mainly determined their passing or failing the class. Besides my test, I also considered the learners' performance on the midterm exam and participation in the classroom.

Reporting scores

Did you hand the test back to learners? Did the learners get to keep the tests? Or did you take the tests back?
As it was a final exam, the tests were not handed back to learners for checking their answers, which I frequently did for the midterm exams. However, the learners could not get to keep the tests as it is a requirement of the university that the tests be taken back and kept by the registrar's office.

Did you offer feedback to individual learners in addition to their test scores? Written? Orally? In or out of class?
I did not offer individual feedback but I shared the answers with learners by sharing the test key electronically as a PDF file.

Did you teach learners to interpret their scores?
I did not teach learners to interpret their scores as the university's student affairs system provided the mean and other basic statistics and how to consider the information in them.

Using test scores
Do you feel you learned what you needed to learn from the test results about what learners could and could not do, or what they knew or did not know?
I actually learned what I needed about my learners. For example, I noticed that they still had problems with complex passive structures although they were provided with a plenty of examples and a great deal of discussion.

Did your test change how learners studied?
My test actually changed how learners studied. Nearly all my learners practiced grammar through multiple-choice questions as the university

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

entrance examinations were based on this type of item format. However, I aimed to encourage learners to create their own responses or sentences based on the given structures. Therefore, my test included a variety of item formats in which my learners had to produce rather than select or recognize the answers.

Did you spend time going over the test in class?
As my test was the final exam, and there was no class after it, I could not spend time going over the test in class. However, I provided the test key to the learners.

Did learners ask you about the test itself (not the test scores) outside of class? If so, what did they want to talk to you about?
Several learners asked me about the test outside of class through emails or face-to-face whenever they and I were available. They mostly wanted to talk about the items whose answers they were not sure about.

Did learners' test scores change your teaching? Did you change your teaching for future courses based on test results?
The item analyses and the test results helped me focus on the structures and areas on which my future learners needed more practice and explanation, and I re-ordered the content depending on their strengths and weaknesses.

If you could turn back time, what would you change about your test? What would you change about your test administration?
If I could turn back time, I would change the dichotomous test items (items with only two possible answers), as several learners explained that sometimes they just selected any answer. This raised concerns about learners just getting answers correct by chance. Moreover, I would also increase the amount of time for learners to take the test since several told me that they could not answer several questions due to lack of time. A great majority of the students, however, said that the test was fair as it covered all the language structures in the lectures and the course content.

Section Four: Evaluating and Reviewing your Answers

To what extent do you think you've described recurrent patterns in your work with tests?
I think I have described recurrent patterns regarding determining test contents, scoring and scoring procedures, feedback provided to learners, and the decisions made based on the test results.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

To what extent is your test here an innovation, or something new, for you?
Regarding the use of a variety of test items which require learners to produce answers rather than recognizing the correct answers, is, to some extent, an innovation for me. This is highly valued by the majority of my learners as they shared their views on my test, saying that alternate test items on every single language structure are helpful for them to put their knowledge into practice. Moreover, using texts related to their own field of study such as the one in the text on dyslexia might also be considered new as I used other types of texts or sentences not directly related to language and language teaching.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

A Speaking Skills Test for High School Learners of English in Southern Tyrol

Gisela Mayr
Free University of Bolzano,
Gymnasium Walther v.d. Vogeweide

Introduction

I am a teacher of English language and culture in secondary education in South Tyrol. South Tyrol is situated in the middle of the Alps and on the border to Austria. It is a well-known tourist destination, particularly for the Dolomites, also called the pale mountains, especially since they were declared a UNESCO World Heritage Site in 2009. However, the region has a troubled recent history which even today influences the languages we currently use, and our language education policy. South Tyrol became part of Italy only in 1923, when previously it had belonged the Austrian-Hungarian Empire and had many German speakers. After the Second World War, South Tyrol was treated by Italy as a linguistic and cultural minority region, and was granted autonomy by the Paris Treaty of 1946. The German-speaking minority was granted an autonomous German school system within the Italian state, where the language of schooling was German primarily, and Italian secondarily.

Italian has mainly spread in urban areas, mostly in Bolzano and Merano, where the Italian-speaking population settled during a period of two decades in the twentieth century. As a consequence, urban areas, particularly Bolzano, are predominantly Italian-speaking, and contrasts with the Germanophone rural areas surrounding it. In addition, in the last decades South Tyrol has become a country of immigration, with people arriving from Afghanistan, Eastern Europe, North Africa (refugees), and Pakistan. This has added even more complexity to a linguistic and political situation that could be characterized as unstable. This context constitutes a big challenge for schools, which need to create learning settings and environments suitable for linguistically and culturally heterogeneous groups.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

In general, South Tyrolean schools are part of the Italian school system, which is inclusive. This means that there are no standardized tests at the end of the school year. Students are supported individually in their learning processes and there are no unitary course objectives. The ultimate goal of schooling is providing educational justice rather than performance according to specific standards. This statutory framework supports cultural and social inclusion of all students, regardless of their linguistic or social background.

The learners taking the test presented here are 14 or 15 years old. Our school is college preparatory and most learners will proceed to university after taking their A-level exams. The students attend the language section of the school and study three foreign languages: Italian, English, French, as well as Latin. The language of schooling is German. This means that also the teachers' first language almost always is German, and this holds for the teachers of English. The teachers of Italian are the only exception to this fact, as they must be Italian native speakers by law. In some cases teachers of Italian might have only an elementary knowledge of German and English. This explains why the scoring sheets in the tests presented here are written in German and Italian.

At the end of their school careers of five years, the students are supposed to have achieved a *CEFR* (Common European Framework of Reference) B2 to C1 level of English and Italian in all five skill areas (Council of Europe, 2001). English plays a predominant role in foreign language education here, as it is considered by many a *lingua franca*. Most universities in Europe use it as an official language of study. Italian, in contrast, is not popular here due to historical reasons, reasons of low social prestige, and also its relative unimportance in economy and science.

The Test

The test format I present is used to assess speaking competencies of students between a B1 and a B2 level of the *CEFR* (Common European Framework of Reference). We use this test at the beginning of the school year as an entrance test for the students who change from middle school (*Mittelschule*) to the Gymnasium (college preparatory school). It can therefore be considered a resource for a needs analysis that helps teachers to plan and organise the aspect of spoken English language instruction during the year. The test gives needed salience and relevance to this aspect of language learning, which is all too often neglected.

The test consists of two parts: a monologic and a dialogic part. The test texts and comprehension questions for the monologic part in English are taken mainly from the magazine *Current* (B2/C1 level) edited by Mary

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Glasgow (https://maryglasgowplus.com/subscribe/english/current). The test texts for the Italian monologic part are taken from the magazine *Focus Junior* (https://www.focusjunior.it/) as well as Italian magazines for teenagers. See Table 4-6 for a short excerpt of a 570-word text and comprehension questions prepared by the author for the text.

Table 4-6: ***Sample text excerpt and comprehension questions***

Sample text excerpt:	
"I write fake news, and there is nothing wrong with what I do." So says an anonymous writer who calls himself Chief Reporter, or „CR". He's the creator of a southeast London website called Southend News Network. CR started writing fake new articles as „a bit of a joke", but also he says as a reaction to how local stories were being covered by the media - some of which bordered on fabrication. „People read a headline and then don't even bother to check the content before they share it," CR says. He then went further and began to simply invent stories on any hot topic: "It ended up with up to two million views per month and many stories ended up on Facebook. What is one of CR's headline news spoofs? SOUTHEND RESIDENTS EVACUATE TOWN OVER SUPERMOON COLLISION FRARS". In other words, CR fooled readers into believing that locals had fled their homes due to the threat of another planet smashing into Earth. The "Supermoon" story is an obvious example of fake news: a type of harmless satire presented to look just like mainstream journalism. In fact satirical news has been around as long as journalism itself.	
Student A comprehension questions	Student B comprehension questions
What is the "Southern News Network"?	What are "fake news"?
Are "fake news" always something bad?	When do "fake news" become too influencial?
What are the standards of mainstream journalism?	Why has information reached an "all-time low" ?
When do "fakes news" become a threat?	Are "fake news" easy to spot? Why not?
Student A personal opinion	Student B personal opinion
Do you think you are a critical reader of news in the internet? Why yes? Why no?	Do you always check the sources when reading news in the internet? Why yes? Why not?
Have you ever spotted any "fake news" ? What did you do?	Do you think "fake" news" have recently influenced the political scene in Italy?
What do we mean by *reading between the lines*? Do you usually try to read *between the lines*?	Do you think that *reading between the lines* might help you to better identify "fake news"? Why?

Note. Adapted and from FAKE NEWS! Are you fooled? Mary Glasgow Magazine *Current,* Issue April/May 2017 Vol. 48. Scholastic, 8-11. Reprinted by permission.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

The test procedure. The tests are taken outside of the classroom. This means that two students at a time are taken out from class and tested. They sit in front of the teachers at a desk. There are two teachers present: One is the examiner and one the rater (see explanation below). First, learners' monologic skills are tested. The students are first asked to summarise a text they were given in advance to read and prepare. See descriptions of test texts above. Then they are asked to answer comprehension questions about the text, and then personal opinion questions (see Table 4-6 above). The students take turns in answering the questions. A sample question for student A for comprehension is "Are 'fake news' always something bad?" and for personal opinion "Do you think you are a critical reader of news in the internet? Why yes? Why no?" A comprehension question for student B is "What are fake news?" and for personal opinion "Do you always check the sources when reading news in the internet? Why yes? Why no?" There are different questions provided for each student in order to avoid repetition, which could easily falsify the results. Recall that learners are assessed in groups of two.

The monologic assessment is aimed at grading the ability of learners, working alone, to summarise a text using their own words and structures, and also to learn and activate new vocabulary and structures. The questions about their personal opinions are meant to check if the students have grasped the text's meaning and ultimately if they are able to express their personal opinion with regard to the topic treated in the test text. This exam format is meant to move form a level of mere reproduction of meaning towards a more cognitively demanding level of critical analysis and evaluation of a text (Bloom, 1976). Therefore the texts provided should be slightly above the B2 level (*CEFR*). The first part of monologic speaking should last an average of five minutes.

The second part of the exam covers the competence area of dialogic speaking. Here the students are supposed to interact and talk to each other in a role-play. The students are given role-cards. See Table 4-7.

Table 4-7: ***Sample role play cards in English and Italian***

English:

Student A: You want to go to the cinema on Saturday, because there are some friends you want to meet. You need someone to come with you. Your best friend is not very outgoing but you have to convince her/him, because you desperately want to be there on Saturday and he/she is the only one available. What arguments can you find to do so?

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Student B: You don't like going out in the evening, especially at weekends - you are tired. You prefer to stay at home and watch TV or invite some friends over. If someone asks you to go out, you usually find some excuse, even if it is your best friend. It is very difficult to convince you but not impossible.

Italian

Alunno/a A: hai visto il film "....." al cinema, ti è piaciuto tantissimo. Pensi che gli attori siano stati bravissimi e che la trama sia stata non solo convincente ma anche avvincente. Hano proprio meritato di vincere l'Oscar. Non riesci a capire come qualcuno non la possa pensare come te. E' un film fantastico. Il/la tuo/a amico/a non la pensa come te, discutetene insieme
Alunno/a B: hai visto di recente il film "....." non ti è piaciuto per niente. Gli attori non erano bravi e la trama scontata e noiosa. Non rieschi a capire perchè abbia vinto così tanti premi. Ti stupisce che sia potuto piacere a tante persone. E' un film che certamente non merita le crtiche positive che ha ricevuto. Il/la tua amico/a non la pensa come te, discutetne insieme.

Sufficient time to read the cards is provided, and where necessary, explanations are given by the teacher. Then the students are asked to act out a dialogue following the instructions given in the role-play cards. This second part of the test lasts approximately another four to five minutes.

Scoring. The scoring is done during the performance and the same procedure is adopted in the monologic and the dialogic speaking phase. In this type of exam one teacher is the examiner and interacts with the students when necessary, putting them at their ease, and the other teacher is the rater who stays also physically behind and fills in the scoring sheets. The rater is given the scoring sheets in advance to analyse them and if necessary, clarifications are provided. The scoring scales are filled in only by the rater during the performance.

On the scoring sheets different competence areas are identified and for each competence area descriptors are allocated (see Table 4-8). There are not more than three descriptors per competence area in order to facilitate observation. For each descriptor, five competence levels are provided, and the rater marks the one he or she deems right. The description of the competence levels is adapted from the school's language unified curriculum, which in turn is adopted and well known by all language teachers for testing, evaluation and assessment. This aids us in achieving consistency of rating for any descriptors used in a class. See Table 4-8 for the competence levels used in the school:

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 4-8: ***Description of competence levels to aid in scoring consistency***

Competence level	Description of the competence level
Is not true	The competency described in the descriptor never occurs. It is not part of the student's active repertoire of speaking competencies.
Is rarely true	The competency described in the descriptor can be observed only once or twice or is present only in outlines. It is not yet an integrated part of the active repertoire of speaking competencies. Frequent errors in the application may occur.
Is sometimes true	The competency described in the descriptor can be observed more often but is not yet a fully integrated part of the student's active repertoire of speaking competencies. Occasional errors in the application still occur.
Is often true	The competency described in the descriptor can be observed regularly and is to a large extent integrated part of the student's active repertoire of speaking competencies. (Almost) no errors occur.
Is always true	The competency described in the descriptor is well known and used strategically in discourse. It belongs to the student's active and extended repertoire of speaking competencies. No errors occur.

The scoring scales. For the monologic test task, learners are scored using the German/Italian score sheet. See Table 4-9.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 4-9: ***Monologic test task score sheet in German and Italian***

Parlato monologico 15.09.15
Name: _____ Klasse:_____ Lehrperson:_____

	Indikatoren	Trifft nicht zu	Trifft selten zu	Trifft teilweise zu	Trifft häufig zu	Trifft immer zu	Bemerkungen
Klarheit von sprachliche Mitteln und Wortschatz	Der Schüler/ die Schülerin verfügt über erweiterte Präsentationskompetenzen im Kontext der gegebenen Aufgabenstellung. L'alunno/l'alunna dispone di competenze allargate di presentazione nell'ambito del compito da svolgere.						
	Der Schüler/die Schülerin kann seine/ihre Darlegung zusammenhängend und Logisch aufgebaut formulieren. L'alunno/l'alunna sa strutturare la propria presentazione in modo logico e coeso.						
	Der Schüler/die Schülerin verfügt über einen angemessenen allgemeinen und fachbezogenen Wortschatz. L'alunno/l'alunna dispone di un lessico generico e specifico appropriato.						
Strukturen	Der Schüler/die Schülerin kann grammatisch korrekt und differenziert sprechen. L'alunno/l'alunna si esprime in modo grammatcalmente corretto e differenziato.						
	Der Schüler/die Schülerin kann komplexe						

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

	Sprachstrukturen anwenden. L'alunno/l'alumna utilizza strutture linguistiche complesse.						
Aussprache Intonation Reflexion	Der Schüler/die Schülerin kann flüssig und klar kommunizieren. L'alunno/l'alunna communica in modo chiaro e fluido.						
	Der Schüler kann Lautung, Akzent und Intonation weitgehend korrekt anwenden. L'alunno/l'alunna utilizza intonazione, accento ed pronuncia in modo corretto.						
Klarheit der inhaltlich- thematischen Darstellung	Der Schüler/die Schülerin kann den eigenen Standpunkt argumentativ geordnet und kohärent darlegen. L'alunno/l'alunna sa esporre il proprio punto di vista in modo ordinato e coeso.						
	Der Schüler/die Schülerin kann das Thema angemessen komplex und differenziert entfalten. L'alunno/l'alunna è in grado di esporre il tema in modo complesso e differenziato.						
Reflexion	Der Schüler/die Schülerin kann die eigenen Fehler selbst korrigieren. L'alunno è in grado di correggere i propri errori.						

For the dialogic test task, learners are scored using the German/Italian score sheet. See Table 4-9.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 4-10: *Dialogic test task score sheet in German and Italian*

Parlato dialogico 15.09.15
Name: _____ Klasse:_____ Lehrperson:_____

	Indikatoren		Trifft selten zu	Trifft teilweise zu	Trifft meistens zu	Trifft immer zu	Bemerkungen
Klarheit von sprachliche Mitteln und Wortschatz	Der Schüler/die Schülerin kann Gesprächstrategien und Redemittel für die Gesprächstrukturierung angemessen anwenden. L' alunno/L'alunna sa strutturare il discorso ed utilizzare gli strunenti del comunicare in modo appropriato.						
	Der Schüler /die Schülerin verwendet einen variablen und aktiven Wortschatz. Kann unbekannte Wörter paraphrasieren und findet Synonyme L'alunno/L'alunna utilizza un lessico differenziato ed attivo. Sa parafrasare parole sconosciute ed utilizza sinonimi.						

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Strukturen	Der Schüler/ die Schülerin verwendet Strukturen korrekt, es kommt zu keinen Missverständnissen. L'alunno/L'alunna utilizza strutture grammaticali in modo corretto, senza impedimenti alla comunicazione.						
	Der Schüler/die Schülerin kann erweiterte und differenzierte grammatische Strukturen angemessen anwenden. L'alunno/L'alunna utilizza strutture grammaticali più complesse e differenziate in modo appropriato.						
Aussprache Intonation Reflexion	Der Schüler/die Schülerin kann Lautung, Akzent und Intonation weitgehen korrekt anwenden. L'alunno/l'alunna utilizza intonazione, accento ed pronuncia in modo corretto.						
	Der Schüler/die Schülerin kann meistens flüssig sprechen, es kommt nur gelegentlich zu Überlegungspausen. L'alunno/l'alunna parla in modo scorrevole, utilizzando poche pause di riflessione.						

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Klarheit der inhaltlich- thematischen Darstellung	Der Schüler/die Schülerin lässt in ihren Äußerungen den thematischen Bezug zu Äußerungen des Partner/in klar erkennen. L'alunno/l'alunna pone chiaramente in relazione le proprie affermazione con quelle dell'interlocutore.						
	Der Schüler/die Schülerin kann das Thema differenziert behandeln und Sachverhalte angemessen und gut strukturiert erläutern. L'alunno/l'alunna tratta il tema in modo differenziato, ed espone fatti in modo appropriato e ben strutturato.						
	Der Schüler/die Schülerin vermag den eigenen inhaltich-thematischen Standpunkt klar darzustellen. L'alunno è in grado di presentare il proprio punto di vista tematico ed contenutistico in modo chiaro.						
Reflexion	Der Schüler/die Schülerin kann die eigenen Fehler selbst korrigieren. L'alunno/l'alunna è in grado di correggere i propri errori.						

See Table 4-11 for a scoring sheet that has been completed by a teacher.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 4-11: ***Completed scoring sheet***

Parlato monologic [illegible] Name: Klasse: Lehrperson:

	Indikatoren	Trifft nicht zu	Trifft selten zu	Trifft teilweise zu	Trifft häufig zu	Trifft immer zu	Bemerkungen
Klarheit von sprachliche Mitteln und Wortschatz	Der Schüler/ die Schülerin verfügt über erweiterte Präsentationskompetenzen im Kontext der gegebenen Aufgabenstellung. L'alunno/l'alunna dispone di competenze allargate di presentazione nell'ambito del compito da svolgere.				X		
	Der Schüler/die Schüler n kann seine/Ihre Darlegung zusammenhängend und logisch aufgebaut formulieren. L'alunno/l'alunna sa strutturare la propria presentazione in modo logico e coeso.				X		
	Der Schüler/die Schülerin verfügt über einen angemessenen allgemeinen und fachbezogenen Wortschatz. L'alunno/l'alunna dispone di un lessico generico e specifico appropriato.				X		
Strukturen	Der Schüler/die Schülerin kann grammatisch korrekt und differenziert sprechen. L'alunno/l'alunna si esprime in modo grammaticalmente corretto e differenziato.			X			
	Der Schüler/die Schülerin kann komplexe Sprachstrukturen anwenden. L'alunno/l'alunna utilizza strutture linguistiche complesse.			X			
Aussprach e Intonation Reflexion	Der Schüler/die Schülerin kann flüssig und klar kommunizieren. L'alunno/l'alunna comunica in modo chiaro e fluido.				X		
	Der Schüler kann Lautung, Akzent und Intonation weitgehend korrekt anwenden. L'alunno/l'alunna utilizza intonazione, accento ed pronuncia in modo corretto.			X			
Klarheit der inhaltlich-thematischen Darstellung	Der Schüler/die Schülerin kann den eigenen Standpunkt argumentativ geordnet und kohärent darlegen. L'alunno/l'alunna sa esporre il proprio punto di vista in modo ordinato e coeso.					X	
	Der Schüler/die Schülerin kann das Thema angemessen komplex und differenziert entfalten. L'alunno/l'alunna è in grado di esporre il tema in modo complesso e differenziato.			X			
Reflexio n	Der Schüler/die Schülerin kann die eigenen Fehler selbst korrigieren. L'alunno è in grado di correggere i propri errori.				X		

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

English translations of the scoring sheets. Presented here are the English translations for the monologic and dialogic test score sheets. See Table 4-12 and Table 4-13.

Table 4-12: ***Monologic test task score sheet (English translation)***

Monologic speaking
Name:_____ Group:_____ Teacher:_____

	Descriptors	is not true	is rarely true	is sometimes true	is often true	is always true	Observations
Clear use of vocabulary and structure	The learner has an extended command of presentation competencies within the context of the assigned task						
	The learner can structure his/her presentation in a logical and cohesive way						
	The learner possesses adequate general and specific vocabulary						
Correct use of structures	The learner is able to express him-/herself in a differentiated and grammatically correct way						
	The learner can makes use of complex grammatical structures						
Pronunciation and intonation	The learner can communicate fluidly and clearly						
	The learner can use accent, intonation and articulation to a large extent correctly						

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Cleanliness of contents and thematic representation	The learner can represent the own point of view clearly and cohesively						
	The learner can expose the topic in a complex and differentiated way						
Reflection	The learner can make use of self-correction during the presentation						

Table 4-13: ***Dialogic test task score sheet (English translation)***

Dialogic speaking
Name: _____ Group: _____
Teacher: _____

	Descriptors	Is not true	Is rarely true	Is sometimes true	Is often true	Is always true	Observations
Correct use of Speaking strategies and	The learner can structure his/her discourse by correctly applying speaking strategies and useful phrases						
	The learner can use a variable and active vocabulary						
	The learner can paraphrase unknown words and find synonyms						
Correct use of grammatical structures	The learner uses grammatical structures correctly and so can avoid misunderstandings						
	The learner can make use of extended and differentiated grammatical structures in an appropriate way						

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Intonation accent articulation	The learner generally can use accent, intonation and articulation in a correct way						
	The learner can speak fluently for the most part, only sometimes hes/she needs to pause and reflect						
Cleanliness of contents and theme	The learner puts his utterance in clear thematic relation to the utterances of his/her interlocutor						
	The learner treats the topic in a differentiated way and presents facts in an appropriate and well structured way						
	The learner is able to clearly present his/her point of view regarding thematic content aspects						
Metalinguistic reflection	The learner makes use of self-correction while speaking						

Contributor's Questionnaire Responses

Section One: Test Planning and Writing

Why did you write the test? What were the purposes of the test?
I developed this test because I wanted the focus of teaching and learning to be more centred on speaking competencies. Usually this competence area is neglected in Foreign Language Teaching (FLT), because the assessment is not very easy. Assessment in general is carried out throughout the school year in Italy and more generally in Europe, and not just at the end of each semester. As a result, traditional interrogations as they are practiced here, are not sufficiently articulated, and there is not enough time to assess speaking competencies. Usually testing spoken language is more concerned about the assessment of content knowledge provided through appropriate language. But no specific tests are carried out where speaking competencies are assessed.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

The test was administered in our English as a Foreign Language (EFL) and Italian as a Second Language (ISL) programs at the beginning of the school year to assess students' level of speaking competencies with regard to monologic and dialogic speaking. Its intent was to help the teacher to better plan the English curriculum for the school year by putting more weight on the development of speaking competencies, as these tests showed where the actual needs were but also where students performed well. This way, areas of intervention were identified for the development of the different levels. On the other hand, the test helped students to better evaluate their own speaking competencies and identify their personal areas of improvement. The test gave them the possibility to speak the foreign language or L2 with a peer for a longer period of time and, as the atmosphere was relaxed, they perceived this as an occasion for practice. No grades were awarded. The test was repeated at the end of the school year to give learners feedback on their progress. It can also be used as an exercise throughout the school year with peer assessment.

How did you decide how many subtests to write? Was there some correspondence between a subtest and a course objective, or was there some other reason to create the subtests you did?

The course objective focuses on spoken competencies in monologic and dialogic speaking, so both areas were assessed with specific test tasks and sub-tasks. In monologic speaking the students had to summarise a text first, and then had to answer specific questions. The first task was cognitively not very demanding (Bloom's Taxonomy, Bloom, 1976) although the language involved was complex. In the second monologic task, the students were asked to reflect on and evaluate the text they had read. This was cognitively more demanding and the language structures required were also very complex. So we can say that the second sub-task was more demanding than the first, although the examiner prepared beforehand the students for the second sub-task by providing examples of vocabulary and structures. Finally, in the last part of the monologic test students were asked to express their personal opinion about the text they had read. This third sub-task can be considered a stage between monologic and dialogic speaking. There is no clear distinction between these two aspects as the boundaries are fluent. In this case the test sub-task required a high register and complex language structures, although students were able to move away from the language they had used previously and use more interactional language. This was possible because, as already stated, this part of the test was no longer clearly monologic, and there was much more interaction between the interviewer and the student than before.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

The role-play on the other hand (the dialogic test task) had the aim to assess everyday and colloquial speaking competencies. What language is needed in informal interactions with peers? What phraseology is present or not? Do students know how to interact? It is quite difficult to practice or simulate everyday situations in class on the B2 level. Special learning settings are needed, which are not usually provided in course books. Authentic input has to be found, but often it is linguistically very demanding and puts weaker students into trouble so they tend not to speak. This kind of test could be a first step to help students overcome their anxiety, and in the meantime become acquainted with more realistic communicative settings.

Did you write as many items or test tasks as you needed, or did you write more than you needed, then pare the number down?
I wrote or prepared more test tasks than I actually used and chose the ones that seemed most suitable, after having discussed them with my colleagues. The topics of the text tasks were chosen so as to be meaningful and interesting for the students. They should be motivating and possibly kindle authentic communication and argumentation. After having developed and collected a sufficient number of tasks, I compared them and discussed them with my co- examiner, and chose those that seemed most interesting for the students, and then matched the requirements of the descriptors best.

The score sheets (Tables 4-9 and 4-10) were revised more than once too. I chose those that descriptors that corresponded best to the B2 level of the *CEFR*. Two other important criteria of choice were that the descriptors had to be observable in a relatively short time (the duration of the assessment) and relevant to the competence area. They had to be clear and understandable for the teachers who assessed the students. Therefore, Therefore, I gave the co-examiner the descriptors in advance so that they could become familiar with them and discuss possible problems with me. Finally, the number of descriptors had to be adapted to the ability of observation of the teachers. The question we had to ask ourselves was: How many aspects can a teacher observe at a time? So not too many descriptors were provided. The chosen number proved to be okay during the first tests, so no change was needed.

How did you decide how many test tasks to write?
The time at disposal for the test was one central criteria in more than one sense, as opposing requirements had to be fulfilled. As the students were taken out of the class and two teachers were needed for the assessment, it

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

had to consume as little time as possible. On the other hand, the test tasks had to provide sufficient time of speaking to render the observation possible and enable the teacher to carry out the assessment. Finally, sufficient time was needed to cover all the competence areas.

Did you have one version of your test, or did you create a second equivalent version?
I had more version, and they were used in the different interviews. This means that not all the students had to read the same texts and to perform the same role-plays. This could be considered negatively, because one could argue that not all the students were given the same opportunity. But as the test scores did not influence students' final grades, this was not so relevant.

On the other hand, it was interesting for the teachers to observe how students coped with different kinds of linguistically complex situations, as each text triggered a different type of conversation. We could observe what strategies the students adopted or needed to handle them and what was actually missing. We integrated this knowledge in our teaching. This was an interesting experience for us, too.

Were you concerned at how long the test would take to administer?
Yes. In fact, time was a central factor in the choice of the length of the test. As the test was carried out outside the classroom and the students were taken out from the lessons two at a time, it was obviously important that the test was not too time-consuming.

Were you concerned at how long the test would take to score?
No, because the teachers had to tick the competence level they wanted to assign in a grid during the testing (Tables 4-9 and 4-10). The results could be scored easily on the computer. This means that the scoring was mainly done during the interview by the teacher who was not taking an active part in the performance test, but was sitting a little behind, observing and filing in the assessment grid.

Were you concerned how you might use the test items themselves for learner feedback?
The students could evaluate their speaking competencies on their own during the test and by looking at their scores and reading the descriptors afterwards. As the scores were discussed together after the interview, they got all the information they needed.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you consider having your students take your test on a computer?
I do not think assessing speaking competencies on the computer would make much sense, unless you want to assess pronunciation, intonation or prosody. We could have recorded the students and then analysed the data, but firstly this would have been far too time-consuming. It is also true that in Italy due to the legal protection of minors, in order to record them, you need parents' permission. It is quite a hassle for the teachers. In this case it would also have given the testing an official note and might have put the students under unnecessary pressure by giving the impression that this is some kind of summative assessment.

Furthermore, students are not used to recording of any type, so I presume this too would have risen their anxiety, which is counterproductive. It also seemed that writing down the score during the interview rendered the process more human, which obviously also implies more possibility of failure. But what is really important here is that there is a relationship and real interaction between the teachers and the students. This way the test becomes part of the learning process and supports the students in further learning. Recording students would have disrupted that.

Did you consider making the test an open book test?
This was in fact and open book exam, as students were given the texts they had to summarise and discuss in advance as well as their role cards for the role-plays.

Did you consider allowing learners to use additional sources such as dictionaries, or their notes, while taking the test?
They could use dictionaries while preparing the texts they had to discuss.

Did you plan to allow learners to re-take a test for improvement? The same test, or a different test?
Yes, it is very useful to let students do this kind of test more than once, because they can observe their own improvement.

What sources did you draw from for your test?
I consulted the following sources.

Common European Framework of Reference for Languages:
www.coe.int/t/dg4/linguistic/source/framework_en.pdf

Mündliche Prüfung im letzten Jahr der Sekundarstufe I – Englisch:
https://www.schulentwicklung.nrw.de/cms/upload/muendl_kompetenzen/

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Muendliche_Pruefungen_in_der_Sekundarstufe_I.pdf 21.10.18
Translation: Oral examination in the last year of grade I - English

Standardsicherung Schulministerium NRW:
https://www.standardsicherung.schulministerium.nrw.de/cms/upload/muendl_kompetenzen/VVzAPO-SI_Anlage_55.pdf 21.10.18
Translation: Quality management of standards Ministry of Education NRW (Nordrhein-Westfalen).

Standardorientierte Unterrichtentwicklung NRW:
http://www.schulentwicklung.nrw.de/cms/upload/ue-englisch/modul_1/teil-3.pdf 21.10.18
Translation: Standard-oriented development of teaching NRW (Nordrhein-Westfalen)

Rahmenrichtliniem für Fremdspracheen Südtirol:
www.bildung.suedtirol.it/.../druckfassung_rahmenrichtlinien_gs-ms-dt09... 21.10.18
Translation: Framework of reference for foreign language teaching, South Tyrol

Journal: *Der Fremdsprachliche Unterricht Englisch.* Issues: 90, 108,133, and 90
Translation: Foreign language teaching English

The idea for the test tasks were taken from tasks we usually do in class. However, the texts used for students to respond to were authentic materials and taken from magazines, newspapers, and online material not offered in textbooks. The role-plays were chosen so as to simulate real situations the students might encounter in the region (real informal situations that might arise between young people), and dealt with problems that are usually not tackled in course books as they referred to realistic situations and places. They were freely invented.

Did you consider learners' communicative competence when writing test items?
Yes, communicative competence was at the centre of interest.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

What aspects of communicative competence?
Competence in general, according to Weinert (Weinert, 2001; see also Hartig, Klieme, & Leutner, 2008) means to tackle and cope with new and demanding situations without previous knowledge. Communicative competence, then, is the ability to communicate using also complex linguistic meanings without fear, to know about their effects on interlocutors and to have a thorough knowledge with regard to the language's meaning and use. This is true even if speaking in a code, which one has not mastered completely and which is only in part present in the personal idiolect (freely translated from Piepho, 1974, 1979). Communicative competence therefore also means to respond to the linguistic ability of the interlocutor and to adjust to it.

Furthemore, the present understanding of communicative competence sees communication not only as the use of language as a system of signs, a code meant to convey meaning, but also as a way to see and interpret the world, symbols, emotions, and expectations. It encompasses the whole personality and body, language puts the individual in contact to other individuals and through this interaction new worlds are created (Kramsch, 2006, 2009, 2011).

If you wrote a performance test (where learners have to converse or present a topic, or write something above the sentence level), how did you get the ideas for what task learners had to do for the test?
I took the idea from tasks students have to do in real life situation in which students might find themselves in: A formal situation e.g. study, university or work, or a life world situation (friends, colleagues etc.). We simulated linguistically difficult situations. The main criteria of choice were as already mentioned were relevance and meaningfulness. The task had to engage learners' interest.

How did you get ideas on how to score learners' performances (the scoring criteria)?
The scoring criteria (Tables 4-9 and 4-10) had to match the criteria of the school's curriculum for language teaching. The school curriculum makes reference to the *Südtiroler Rahmenrichtlinien für den Fremdsprachenunterricht* (South Tyrolean Framework for Foreign Language Teaching, http://www.provinz.bz.it/bildung-sprache/deutschsprachige-schule/default.asp) which is a statutory provision passed by the local authority and is mandatory for all schools. The criteria for this framework as well as the descriptors contained in the document were derived from the *CEFR* and adapted to local requirements.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

If you used points on a scale for scoring learners' performances, how did you decide on how many points on a scale to use?
The scales needed to cover a differentiated spectrum of levels so to give the teacher or examiner carry out accurate scoring of individual students, taking into account even little differences. On the other hand, the scales had to be diagnostic and therefore it was important that scale have too many points on it.

Were you concerned about whether you could get a colleague to help you score learners' task performances?
For this test it was necessary that one teacher do the testing and the other do the scoring. So the test needed two teachers. But I was not concerned. My colleagues were all very helpful as it was part of the additional working duties that they were paid for.

Did you plan to give the scoring criteria to the learners for future self- or peer-assessment?
As it was an open book assessment, students had access to the score criteria (which in this case correspond to the descriptors), but only after the test. Peer-assessment can be carried out during the school year as an exercise.

Did you seek help from a peer to clarify what your test items or tasks were measuring?
I got feedback form colleagues as well as a university professor who helped me. It was in fact important that the descriptors met the actual tasks I had prepared. So, I compared the descriptors with the texts I had prepared students had to respond to, and the tasks the students had to do during the test. I took care that the requirements matched. For this I revised the descriptors more than once.

Did you compare your test to the lessons that learners had?
Most of the time, speaking exercises students practice in lessons are not very realistic, because they mainly follow the textbook instructions. This renders them artificial as the topics treated are carefully selected so to respect the variety of religions and cultures around the world. One can say that they are culture and religion neutral. Even if the books are relatively new, they are already obsolete and do not reflect the actual state of current life. When developing this test, it was my aim to bring the students face to face with authentic input on current issues and topics. They should be really interested in what they were reading and talking about, and this way

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

by being really engaged in what they were saying and discussing, move to a more realistic form of communication than is usually possible in school.

Did you compare your test to the textbook or other materials learners used?
One of the aims of the test was to amend the textbook tasks with real world tasks, and where possible to add new competencies that better reflect real life requirements. Textbook tasks are in many cases not sufficient to make learners fit for real world communication, as they seldom provide tasks where students are really involved in the topic and authentic communication can take place.

Adapting existing tests

Are you required to use specific tests in your program?
We are not required to use specific tests, however the tests have to cover all the five skills (reading, writing, listening, and monologic and dialogic speaking). One of the newer competences proposed by *CEFR*, mediation, is not yet stipulated. The *CEFR /CV* was published only in 2018 (Council of Europe, 2018) and is not yet contemplated in the local framework for language teaching and learning.

Was your test different from required tests? How?
It did not differ substantially from required tests as there is no such thing, but it differed form usual tests. As already explained at the beginning, it was a new way of testing, which differed substantially from habitual testing practices at the school.

Section Two: Test Administering and Scoring

Were you concerned about test security?
As it was a spoken interview and only two students were present outside the class, I was not concerned about test security.

Were your tests photocopied, or did learners see your test items or test tasks another way, such as on the blackboard?
The score sheets were photocopied for the teachers to fill in. They got one copy for each student. The instructions (texts, questions and role-play cards) were photocopied for the learners. They could hand the instructions on to the next students taking the test, and prepare for the interview.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

How did you deal with learners who missed the test, or who were late for the test?
Failing in this case was not possible as no grades were given. Competencies cannot be graded negatively, you can only decide whether they are present or not and to what degree they are present. Absent students were given another chance. In this case this was quite difficult, because the two teachers had to be present yet again. So, we assigned all the absent students the same date, when they had to present themselves for the interview.

How did you prepare learners to take the test? Were there any test items or test procedures that were new to them?
The test tasks themselves were not completely new to the students, as we used similar tasks in class. It was, however, a completely new and unexpected testing procedure to them and the text content was also much more recent than that usually found in textbooks. The real challenge, however, was the procedure. So, to make it easier for the students, they could prepare themselves thanks to the instructions and texts provided beforehand for reading. The test procedure was clearly explained to them, together with the scoring, the purpose of the scoring, and the aim of the test. So, the students knew well what was expected of them and they were only a little excited but not scared. This was important, because especially in speaking assessment emotional factors play an important role and anxiety should be reduced as much as possible.

Did you pilot your test? Do a trial run? Did the pilot result in any changes to the final version of your test?
The first time was actually a trial run but no major difficulties emerged, so the instructions for the students and the tests were not changed. As far as the score sheet is concerned, the Italian translation of some descriptors had to be reformulated to eliminate ambiguities for the Italian teachers.

Did you write any of the test in the learners' first language?
No. I wrote the score sheets in two languages for the teachers, as Italian as well as English were assessed. The tests were administered only in the foreign languages tested.

For performance tests: How did you accomplish scoring learners' performances?
As there were two teachers present, it was possible to score during the performance, and give each other a short feedback after the interview (real

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

time scoring). We did not interview more than eight to ten students per hour at a time and took a break after one hour, because it was very fatiguing. There were always two students present for the test, in order to make the dialogic speaking test task possible and to compare their performances.

Real time scoring has the advantage that it is done at the moment when all teachers' perceptions are still very vivid and present. There are aspects such as non-verbal communication that cannot be assessed as a stand-alone feature, but they are a basic element of communicative competence. Therefore, in this case it was better to assess the situation as a whole taking into consideration all components, instead of picking out specific aspects and analysing them into depth. To summarise: The scoring was real time scoring, done by two teachers. Although one was the official scorer, and I was the interviewer, and at the end of each test we discussed the scores briefly and so compared our assessment.

Do you think your test was reliable? What did you do to check?
Reliability in a test of this kind is very hard to achieve. Scoring is always subjective. The things I did to make the test as reliable as possible was to set a number of criteria for the choice of the teachers who did the actual scoring on the sheet. I chose teachers for scoring who did not know the students they were assessing. This way they would not have any preconceptions. The scorers had to be proficient speakers, though not necessarily native speakers, who were very experienced, had a positive attitude towards students in general. Finally, I made sure to be able to keep the same teachers for the whole testing procedure. I, on the other hand, knew many of the students, but I had no direct influence on the scoring. But it was interesting to compare my idea of individual students with the scoring given by my colleagues.

The fact that the scores were discussed, and the different impressions compared between the teachers, also helped looking at results from different perspectives, making the procedure more reliable. I put lot of effort in preparing the scoring sheet and choosing appropriate test texts and test tasks. I discussed the descriptors and the test items with different people and used their feedback to improve the test. As this is the first time such a test is carried out at our school I had to check it over and over again. One can say that I did everything I could while preparing the tests to make them as reliable as possible. However I did not check reliability afterwards.

The conditions under which the students took the tests were similar though not the same. It is almost impossible to replicate an identical

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

speaking test dozens of times in a row. Even little differences can alter the results. I did what I could to give equal test experiences to all students. Therefore, it was important for me to have the same attitude and behaviour while interacting with the different pairs of students, in order to grant them similar conditions. We also tried to carry out the tests in a relatively short period of time, which was within approximately two days for one class. In sum, two classes were tested in two languages, English and Italian. This should give the students the possibility be involved with something new and interesting. If the testing time had taken too long, it would have had a tiring effect on the concentration of teachers and students likewise.

As the texts and the role-plays were not the same from student pair to student pair, one could argue that validity was not assured. However, it was our aim to learn something about students' learning processes. Moreover, the descriptors in the scoring sheet refer to general competencies that spread over a wide range of areas and meanings, and can be applied to different test tasks.

Reporting scores

Did you report the scores to learners?
We briefly discussed the scores with the learners after the interview, so they could compare our scoring with their own perception of the interview.

What was your goal in reporting the test scores to learners?
Students learnt about their strengths and weaknesses in speaking. This gave them the opportunity to regulate their own learning processes during the school year and to focus on the competence areas they wanted to improve. At the same time, the areas where they had performed well were highlighted. Generally, it can be said that this type of testing, thanks to the long speaking time allowed to the students, gave them the impression of being proficient speakers and boosted their confidence. This turned out to have a positive backwash on further testing and speaking.

Did you teach learners how to interpret their test scores?
We discussed the scores at the end of the interview so no special training was necessary. I don't think that for this kind of test it would have been useful to show the students the descriptors in advance and explain them in detail. They were given general information about the procedures. Giving them the descriptors could easily have inhibited their speaking (the caterpillar who thinks about moving all his legs). As it can be considered an entry test and no grades were awarded, there was also no need for this.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

If it had been a graded test, then it would have been correct to show them the descriptors. As we only wanted to analyse the *status quo*, not showing them the descriptors made the situation more authentic and therefore more informative for us.

How quickly did you report scores to learners?
In this case the report was given immediately after the test.

Section Three: Using Test Scores

Cut scores

How did you decide which scores were passing or failing scores (cut scores)? How did you decide which scores meant a specific grade on a test?
In this case no grades were given to the students, as their competence levels were assessed for diagnostic purposes, and their scores would not influence the final grades of the school year. The assessment of students' competencies differs substantially form other types of assessment which imply grading and scoring, because it can only be determine to what degree a competence is present, but you cannot determine the degree of absence of a competence. The competence level cannot be seen as a score and therefore no grades are awarded.

How did you use learners' scores from this test?
The scores were used only by the teachers who took part in the project. The tests and scores were part of the learning process and had the aim to support the teachers when planning future lessons. The aim was also to support the students to know more about their own language proficiency. Another aim was to make students more acquainted with still unexplored areas of language acquisition.

The scores had no institutional consequence for me as a teacher, nor did they have any immediate influence in the institution. However, as many colleagues knew about this project and some of them were already involved the test and the resulting scores might raise others' awarenesses and change the prevailing way of handling and testing speaking competencies.

What was the role of the test score in determining learners' grades?
The test scores did not influence learner's grades, as it can be considered an entry-test for spoken communicative competence. But nevertheless,

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

students' scores influenced the way we handled speaking competencies in following class meetings. They changed our teaching focus, and so the test and the scores might have influenced learners' grades indirectly.

Were other measures used to decide learners' grades, besides your test?
Yes, during the school year many other tests are administered to the students with different measures, and grades are given. The test had a big weight for me, as it influenced the way I taught. It influenced the way I formally assessed speaking competencies during the year, because thanks to the descriptors I had become a more careful listener.

What was the relationship of the other measures to your test?
The other measures belong to the type of testing generally called "summative assessment." It is not their function to give students feedback on their learning, or they do so only implicitly. My test belonged to the kind called "performative assessment" and its function was to help the students better evaluate their learning process and identify areas of possible improvement, without influencing their final grades. This in fact was one of the positive side effects, that learners could assess their own proficiency level, without the fear of their final grades being influenced by that.

Did your test capture some knowledge, skill, or ability the other measures did not capture?
It was particularly focused on life, the world, communication, and communicative interaction, not taking into consideration content knowledge. The test reflected a picture of what authentic communication could be like even at school, and how students cope in situations where the traditional structures given by textbooks are not present. It is an opportunity for classroom discourse to open up in the direction of life with its complexity and ambiguities. In the test, teacher-student interaction is substituted by student-student interaction where answers to questions are not already known, and there is more than an intention to have a mere language/grammar exercise. The language is used to discover something new, and to tell others about some realistic content they do not already know, to surprise them. In only this way can listening and speaking become truly engaging, and can communication acquire a scope beyond that of practicing grammar or repeating knowledge.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Reporting scores

How did you report scores to learners? Was timeliness of concern to you?
Shortly after the interview. This was the best time, so that all memories and impressions were still fresh and could be discussed.

For performance tests, did you use the test criteria to help learners interpret their scores?
The descriptors were in fact criteria. They were shown to the students at the end of the test, and their scores were explained to them accordingly and discussed together.

Using test scores

Do you feel you learned what you needed to learn from the test results about what learners could and could not do, or what they knew or did not know?
It was the aim of this test to adjust my teaching to the needs of the students and to collect all the necessary information to do so.

Did your test change how learners studied?
I hope so, as the test switched the focus of teaching and learning at school. It may have influenced also the way students learn. However, I had no possibility to check on his.

Did you use particular item types or a performance test to change learners' practices or support their learning? Did their scores indicate they had changed their learning practices?
The test was intended to support learners' learning practices. As it should help them to identify what they really needed to become competent speakers not only inside the classroom, when for example content is tested, but also outside classrooms in complex communicative situations. I think learners become acquainted with what is means to use a language in every day situations, and what specific vocabulary and structures are required, even thought the test is only a simulation! But thanks to the current topics in the test texts, students got involved and authentic communication took place. That said, the scores were not meant to show changes in student's learning, but I think they were directed towards the future. This should induce learners to change the way they learn if necessary.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you spend time going over the test in class?
No, the test was discussed only individually.

Did you change your teaching for future courses based on test results?
The aim of this test was to identify areas of improvement for further teaching. Therefore, my teaching was changed in so far as I tried to practice more everyday spoken language, which is not provided in most textbooks. I introduced more current affairs from newspapers and other media in order to stir authentic discussion in the group. In these units I provided the necessary language step by step in a language loop, when needed, and let the students work autonomously.

This caused also my role as a teacher to change: I became more a coach. There was a shift in perspective. The language students studied became attuned to content, and not vice-versa as it usually happens in language classes. I skipped all the content in the textbook that seemed no longer relevant to me, nor to students.

I did not change the amount of homework students did, but I introduced more authentic reading of texts and books, so the discussion in class could be lively.

If you could turn back time, what would you change about your test? What would you change about your test administration?
Nothing, I consider it actually quite useful as it is.

Did learners give you feedback on the test? Did they think the test was fair, or helpful?
The students liked the test and gave a positive feedback. They said it had given them the opportunity to use the language autonomously for an extended task, so they saw it as a good opportunity for practice. They got the impression of being competent speakers, so the testing supported them in their learning. They were not afraid of speaking, because they knew that the score would not influence their final grades.

Did others (parents or administrators or colleagues) give you feedback on the test?
The colleagues gave me a feedback, saying that it was a very interesting form of test, but too time consuming.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

If you used the test and test scores for additional learning opportunities, did anything about that process help you revise the test for future use?
The test was intended to initiate a number of changes in teaching and learning, but obviously it could be improved. It would be good to try and do the test in class as an exercise using peer assessment and in the meantime the teacher could go round, observe, and take notes or score. This way, by using a more informal surrounding, authentic communication could be triggered and the group take over a scaffolding function for weaker learners, who could profit more from social- and imitation-learning. Meanwhile the learning process could be observed throughout the school year, while the students practice speaking, which is what they like best.

Section Four: Evaluating and Reviewing your Answers

To what extent do you think you've described recurrent patterns in your work with tests?
The recurrent task patterns are: Summarising a text, answering questions about a text, expressing one's own opinion, and doing role-plays. This is something I often test also in traditional interrogation-type tasks. Knowledge of a text is tested because it is a necessary requirement for any further conversation and reflection. You cannot discuss a text that you do not know about. Also, the role-plays were not new for the students. They had been used as exercises in class.

What was new was the time devoted to these different aspects of speaking, the focus on the language, and the way the test was administered. The focus on the language has to be highlighted here. As there was no need to assess any kind of knowledge, we could use the time and concentrate on speaking competencies. So it wasn't the testing which was completely new but rather the way it was assessed.

I do not believe in testing being something completely objective and it is never the conclusion of a process (this might be true for summative testing). I believe tests are more the beginning of something. Therefore, testing at school should always be part of the learning process, otherwise it does not make much sense. I perceive testing as part of teaching and think it only makes sense if it supports learning. Students should not be afraid of tests. Students can accomplish great things, if only you give them the opportunity and the space to do so.

A focus on grammar is still considered central by many teachers. This should change. As the CLIL (Contents and Language Integrated Learning) experience shows, there is much more, and deeper, learning when the

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

language is functional to content. This should also be reflected in the way we teach. Learning by heart and reproducing should be substituted by authentic problem-solving where students are engaged in real communication and in groups work on their own solutions and output. When testing I try to support these processes and induce students to think autonomously and also to reflect their learning.

To what extent is your test here an innovation, or something new, for you?
It was the first time I carried out such a test where the students were taken out from class, and another teacher was the main scorer. It was also the first time at our school, and I do not think this kind of test has been carried out even on a national or European level. Similar tests are administered during the Cambridge Certificate exams, however students' speaking is recorded and analysed later.

It was also the first time I used a competence grid for assessment speaking competencies, though I do use such grids for presentations and give the students the descriptors and criteria in advance as this is helpful to them when preparing the presentation. The descriptors made the scoring more precise, and it also schooled my ability of observation. I realised how multifaceted speaking actually is, and how much these various aspects of speaking influence our perception of the whole.

It was also the first time I prepared speaking assessment in such detail. This helped me to better identify the different competence areas expressed by the descriptors. I became a more analytical listener.

Doing assessment with another teacher was new to me, too. I am acquainted with this kind of assessment from doing final examinations, or doing examinations at a university. But scoring the same test together with a colleague teaching the same subject was new. It showed me that sharing and exchanging opinions about how to score students gives a much broader view of the whole. By sharing views, the picture you get is more complete. You are forced to overcome unconscious constraints that are put on you without you even knowing.

It was also new for me to assess speaking as such, without being concerned with contents or knowledge of any type. This way it was possible to closely analyse the students' speaking performance, and to identify aspects which had been overshadowed by the need to check knowledge as you might find in typical school assessments. When planning lessons, I could now take these aspects into consideration and focus on them with special exercises.

It was also interesting to see how differently the teachers tend to score during the test. There was, for example, a rather surprising difference

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

between Italian as a Second Language and English as a Foreign language teachers. It seemed to me that the Italian teachers tended to be stricter. Maybe it was because the Italian teachers were native speakers and perceived spoken language differently. This showed me that scoring, especially with speaking competencies, is always subject to the influence of many different factors, which cannot be completely controlled by the examiner. Having the pretence to be an objective scorer becomes more and more an ideal, which cannot be achieved.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

AN ENGLISH COLLOCATION KNOWLEDGE TEST FOR COLLEGE-LEVEL LEARNERS AND PRE- AND IN-SERVICE TEACHERS

SAKAE ONODA
JUNTENDO UNIVERSITY

Introduction

I am a Professor of English Education in the Faculty of International Liberal Arts, Juntendo University, Japan. I have over 18 years' experience of English teacher training as well as extensive experience of teaching English in high schools. In recent years, I have felt that assessment plays a critical role in language teaching and learning. To examine the teaching methods and techniques I demonstrate in the English teacher education seminars I lead, I have conducted research in assessment methods that can be easily and effectively used by high school and university English teachers in Japan. The test I present here is a product of this research.

This test derives from the Japanese Ministry of Education's *Course of Study* and can be used as a diagnostic test to measure students' knowledge of English verb collocations. Since this is a written test of verbs that most frequently collocate with particular words (though in some cases, two answers are possible), the test is easy for teachers and students to score. If the teacher instructs pairs of students to score each other's answers, it is easy to calculate scores, calculate total scores, and thus determine how easy or difficult a particular item is for students. Thus, this test can be easily used by teachers in senior high schools and universities in terms of content, construct, and format.

This test can be effectively used with senior high school students, university students, and pre- and in-service teachers with a range of English proficiency levels from A1 to C1 on the Common European Framework of Reference (Council of Europe, 2001). The test can be used to both diagnose students' knowledge of collocations and raise their awareness of their importance. I think this awareness raising is a critical

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

factor in making learners' spoken and written production sound fluent and natural. Generally speaking, Japanese EFL (English as a Foreign Language) learners, especially those with low to intermediate English proficiency, tend to transfer collocational knowledge in their L1 (Japanese) to L2 English collocations. As a result, they may convey unintended meaning, or even no meaning at all, or their utterances may not sound natural even if their meaning is understood.

This test (or activity) can be used with senior high school students to diagnose their collocational knowledge at the beginning of a course or to investigate improvement in such knowledge at the end of the course. The test is particularly useful for commonly taught "English Communication" or "English Expression" courses in which the four language skills should be integrated with a primary focus on spoken and written production.

This test can also be used with university students taking a broad range of English courses, including content-based courses in which the primary focus is placed on learners deepening their knowledge, generating opinions, and exchanging such opinions in class by using English. However, it is often reported that these courses do not provide learners with sufficient opportunities to reflect on their language use, including errors caused by L1 transfer, which may persist unless these students are given corrective feedback on their language use.

Finally, this test can be used in pre-service undergraduate English teacher education courses and in-service English teacher seminars in order to raise students' awareness of the importance of collocational knowledge in improving accuracy and fluency in spoken and written production. English teachers will benefit greatly from this test. They themselves may be weak with English collocations, a deficiency often reflected in their own language use. Unfortunately, this sets up a dynamic whereby they rarely or never teach collocations in class. Given that tests of this type are rarely used in Japanese EFL contexts, this particular test will provide teachers and students with new perspectives on teaching and learning grammar and vocabulary as well as speaking and writing.

In my own research, I used the test with first-year university students taking "Interactive International English" and second-year students taking "English for Global Citizenship," both of which are content-based courses with, as stated above, a primary focus on students deepening their knowledge, generating opinions, and exchanging opinions using English.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

The Test

Target audience: High school students; University students; Pre-service and In-service teachers

Target measurement skills:

1. Knowledge of common verb collocations
2. Knowledge of multi-word units

English Verb Collocation Knowledge Test

Directions: Fill in each blank with a word that best suits the context so that each sentence makes sense.

1. I (_____) judo when I was an elementary school student.

2. When she left Japan to study in the U.S., my girlfriend said: "I'll (_____) you."

3. She (_____) me a joke, but it wasn't funny at all.

4. You can (_____) a difference if you are determined to pursue your ambition.

5. You need to think carefully before you (_____) action.

6. It is a Japanese tradition for people to (_____) miso soup with chopsticks.

7. I hear Shota will (_____) 15 on July 1st, and we are thinking of holding a surprise party for him in class.

8. English will (_____) naturally to you if you work hard at it because practice will help you improve your English skills.

9. I believe that being trilingual will (_____) a long way in helping you conduct business in the global market.

10. When we left the restaurant, Dorothy said to me with a charming smile: "Let's (_____) in touch, OK?"

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Answer Key of the Test

Directions: Fill in each blank with a word that best suits the context so that each sentence makes sense.

1. I (practiced) judo when I was an elementary school student. (Partial-credit answer: practice)

2. When she left Japan to study in the U.S., my girlfriend said: "I'll (miss) you." (Partial-credit answer: mis)

3. She (told) me a joke, but it wasn't funny at all. (Partial-credit answer: tell)

4. You can (make) a difference if you are determined to pursue your ambition. (Partial-credit answer: made)

5. You need to think carefully before you (take) action. (Partial-credit answer: took)

6. It is a Japanese tradition for people to (eat) miso soup with chopsticks. (Partial-credit answer: ate)

7. I hear Shota will (turn/be) 15 on July 1st, and we are thinking of holding a surprise party for him in class. (Partial-credit answer: become)

8. English will (come) naturally to you if you work hard at it because practice will help you improve your English skills. (Partial-credit answer: become)

9. I believe that being trilingual will (go) a long way in helping you conduct business in the global market. (Partial credit answer: help)

10. When we left the restaurant, Dorothy said to me with a charming smile: "Let's (keep/stay) in touch, OK?" (Partial-credit answer: be)

Test/Activity Procedure

After students take this test, the teacher asks students to form pairs and exchange test sheets. Then, the teacher elicits the answer to each question from the students and checks whether the answer is correct or not.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Through this procedure, the teacher can give oral corrective feedback to help students internalize knowledge, find other possible answers, and amend answers that got only partial credit (e.g., "told" for "tell" or vice versa). This helps students score each other's test sheet as accurately as possible. Then the teacher has the students calculate the total score, write (usually encouraging) comments about their partner's performance, and return the test sheet to the test-taker. Finally, the teacher collects the scored test sheets and checks for accuracy in the scoring because students who are weak in English may not always score correctly. After the test, the teacher can explain how students can improve their knowledge of collocations.

Contributor's Questionnaire Responses

Section One: Test Planning and Writing

Why did you write the test? What were the purposes of the test?
I had three main purposes in mind when writing the test: 1) To raise learners' awareness of English collocations, i.e., what words are often used in combination with high-frequency verbs; 2) To improve students' spoken and written language production and making it more fluent and accurate by automatizing such fundamental language units; and 3) To minimize learners' adoption of L1 (Japanese) transfer when expressing their thoughts and feelings in the L2.

The test presented here reflects one of the objectives of the courses I teach, including Interactive International English (Year 1) and English for Global Citizenship (Year 2) in the Faculty of International Liberal Arts, Juntendo University. Students' TOEFL ITP scores (https://www.ets.org/toefl_itp) range from 380 to 627. In these courses, the primary focus is placed on production activities such as discussions, presentations, debates, and essay writing. The objective is to help learners interact through intelligible, natural spoken communication and to write reader-friendly essays. For both purposes, it is important to improve accuracy and fluency in learners' spoken and written production.

How did you decide how many subtests to write?
I decided to make one subtest with multiple versions (only one version is presented here). I based my tests on my own error analyses of learners' spoken production as observed in discussion and presentation tasks as well as written production reflected in learners' essay writing.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you write as many items or test tasks as you needed, or did you write more than you needed, then pare the number down? How did you decide which items or test tasks to keep? How did you decide which items or test tasks to discard?

I wrote more items than I needed and then selected those that worked best for measuring learners' knowledge of verb collocations. I selected items that discriminated learners with good knowledge from those with poor knowledge by applying Rasch measurement analyses (Linacre, 2014) to the learners' responses to the items by conducting pilot testing with 20 items. If there was not enough time to conduct pilot testing, I asked a focus group of students who were all pre-service English teachers with high English proficiency to check whether some items were difficult or easy compared with the rest of items and whether there were enough contextual clues for them to understand the sentences and situations. Based on these results, as well as my own intuitions built up over 30 years of teaching, I then discarded items that did not discriminate effectively between learners with different levels of ability with collocations.

How did you decide how many items to write for each test form?

Since I decided to make more items than I needed and to select the best ones, I tried to write twice as many items as I needed to make multiple versions of the test. In addition, I considered the reliability, validity, and practicality of the tests, even though in practice, it was difficult to satisfy all three criteria. I also considered the number of class meetings available because teachers cannot give a test in each class meeting, thereby sacrificing time to be spent on normally scheduled content-based teaching.

Did you have one version of your test, or did you create a second equivalent version?

I created multiple versions of the test by piloting it with groups of students who demonstrated similar learner characteristics as the target students in terms of nationality, proficiency, and age.

Were you concerned at how long the test would take to administer?

Yes, definitely. Since regularly scheduled content-based class teaching must be given priority and learners' concentration does not typically last more than 30 minutes or so, I tried to create tests that could be completed within 10 to 15 minutes.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Were you concerned at how long the test would take to score?
Yes, for two main reasons. First, Japanese university teachers are required by the administration to teach many English classes and they cannot afford to spend too much time scoring tests. Second, it is important to inform the learners (i.e., the test-takers) of their test results so that they can find out what areas they were good at and where they should improve. Therefore, a limited number of items with different difficulty levels will offer students and teachers rich information about students' knowledge of collocations and insights into pedagogical endeavors the teacher might undertake.

Were you concerned how you might use the test items themselves for learner feedback?
Yes, definitely. One of the main purposes of administering the test is to provide learners with a diagnosis of their language development in terms of course objectives.

Did you consider having your students take your test on a computer? Why or why not? If not, what were your concerns?
Yes, I did consider online testing because it would reduce my workload in scoring the test and also give immediate feedback to learners. However, I have misgivings regarding online testing because I am not sure I could program the test in such a way as to measure the learners' partial knowledge and give partial credit for it.

Examples of partial credit decisions are:

5. You need to think carefully before you (take) action. ("take" is a full credit answer, and "took" is a partial-credit answer – correct verb but incorrect verb tense)

6. It is a Japanese tradition for people to (eat) miso soup with chopsticks. ("eat" is a full credit answer and "ate" is a partial-credit answer – correct verb but incorrect verb tense or spelling error)

Did you consider making the test an open-book test?
No, because the test was designed to measure learners' automatized language knowledge.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you consider allowing learners to use additional sources such as dictionaries or their notes while taking the test?
No, because the test was designed to measure learners' automatized language knowledge.

Did you plan to allow learners to re-take a test for improvement? The same test, or a different test?
I did not allow learners to re-take the same test because Japanese students tend to memorize answers without understanding the target concept. For this reason, I planned to give a different test.

What sources did you draw from for your test items?
Test items were drawn from the course objectives as well as what learners do in class and for homework, i.e., both their spoken and written performances. I considered learners' communicative competence when I wrote the test, especially their grammatical competence.

Which test item formats do you prefer to use?
I prefer short-answer or fill-in-the blanks tests for practical reasons.

What types of learner knowledge do you believe you are capturing in your test? How does that change with test item types you used on the test?
I aim to measure learners' knowledge of grammatical verb collocations and multiword units. If I used a performance test or a test that requires long answers, learners with lower or lower-intermediate proficiency may be unable to answer. As a result, it may be difficult to measure these learners' knowledge accurately.

Did you seek help from a peer to clarify what your test items or tasks were measuring?
I checked with a colleague to confirm what I tried to measure, i.e., knowledge of verb collocations and multi-word units. My colleague agreed with my rationale by checking the test items.

Did you compare your test to the lessons that learners had?
I tried to use language items the learners had learned, been exposed to, or used in lessons.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Adapting tests

Are you required to use specific tests in your program?
Yes, we have to use TOEFL ITP (Test of English as a Foreign Language - Institutional Testing Program: Council on International Educational Exchange, 2019), presentations, and essay writing.

Was your test different from required tests? How?
Yes, very much so. TOEFL ITP measures test-takers' knowledge of grammar and written expressions in multiple-choice testing format by asking them to choose the correct option out of four on offer. They must fill in a blank with the one option that is grammatically correct in a complete sentence. Also, the sentences used for the test include many difficult academic words and complicated structures. In contrast, my verb collocation test asks learners to write in the correct verb in the appropriate conjugation to collocate with a particular word in the sentence.

Did you inherit your test or parts of your test?
Yes, in the sense that I copied the format of the question items in Cambridge English exams (University of Cambridge Local Examinations Syndicate, 2019) and adapted them to focus on collocational verb knowledge. I also noted that the structure and written expression section in TOEFL ITP includes sentences with many difficult academic words and complicated structures, and as a result, they cannot be used as they are. Thus, I changed the wording of the target items in my test to make them more familiar to the learners and to make the sentences simpler.

Section Two: Test Administering and Scoring

Were you concerned about test security? What did you do to ensure test security?
Given that I teach three classes in the same course, I was very much concerned that the test items might be leaked to the learners in the other classes that were given the same test later. Therefore, in order to ensure test security, I created several different versions and randomized them to use with particular classes.

Were your tests photocopied, or did learners see your test items or test tasks another way, such as on the blackboard?
The test sheets were distributed to students and collected after they took the test and scored each other's answers in pairs.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

How did you deal with learners who missed the test, or who were late for the test?
They could not take the test as a make-up because they had been told not to be late or absent on test day.

How did you prepare learners to take the test?
I made an announcement about the test format, date, duration, and number of questions a few days before test day.

Did you pilot your test? Do a trial run? Did the pilot result in any changes to the final version of your test?
I piloted the test with different Japanese learners with similar proficiency, motivational levels, and ages meeting as a focus group. Based on the results and feedback from the focus group, I changed sentences in which verb collocates were used but verbs were collocated with different complements. For example, if I received feedback such as "I can't imagine the situation clearly from this sentence" (because the situation is culturally unfamiliar or the respondent has never encountered such a situation), or "There is not enough information for me to understand the sentence" (because of a lack of contextual information), or "I don't understand a word used in the sentence" (because of a lack of knowledge of a particular vocabulary item)," I worked with the focus group to revise the sentences, replace difficult words with more familiar words, add more information, or in extreme cases, change the sentence itself. Also, if nobody could answer a particular question, I changed the target collocation with an item they were more familiar with. On the other hand, if they gave positive feedback to a question, I would leave it as it was.

Did you write any of the test in the learners' first language? Why?
I did not because as I describe above, I made every effort to make the sentences familiar and easy to the learners.

Was your test administered on a computer?
I did not use a computer to administer the test because computer rooms were not available.

How did you get ideas on how to score learners' performances (the scoring criteria)?
I use a set of criteria in which full credit is given for a correct answer, and half credit for an answer with incorrect verb tense or spelling, as in correct answer = 2 points, correct answer but with tense or spelling error or the

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

choice of a verb that rarely collocates with the word that follows but helps convey the meaning = 1 point.

Were you concerned about whether you could get a colleague to help you score learners' performances?
I considered showing the set of criteria discussed above, which I consider clear-cut. But I was still somewhat concerned about asking a colleague to help me score learners' performance as there may be minor differences in interpreting the criteria, such as spelling errors. Likewise, if I score learners' answers on two different occasions, I cannot guarantee that I will treat ambiguous answers consistently.

Did you change your plans or your test to reflect time constraints in terms of test scoring?
I did not change my test because of time constraints because it was fairly easy and not particularly time-consuming to score.

Did you plan to give the scoring criteria to the learners for future self- or peer-assessment? Did you make another, perhaps simpler or shorter, version of the scoring criteria for learners to use?
I gave the learners the scoring criteria because it is a good idea for pairs of learners to score each other's answers as this practice helps them check their internalized knowledge. Of course, as the teacher, I checked their answers later to confirm accuracy in their scoring.

For classroom tests: How did you accomplish scoring learners' tests? Did you score learners' tests twice for accuracy? Did you write a test key? Did you consider alternate answers and add them to the key?
I scored the tests once. I wrote a test key including other possible answers.

Did you hide learners' names as you scored?
That's a good idea, but I did not.

Did you go back and change your marks on previously scored tests in response to problems you found while scoring tests later in the process?
I often do, but given that most the questions on this test had only one possible answer, I did not encounter any problems, so I did not do so in this case.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you put learners' responses to items into a spreadsheet for further analysis? Did that process help you catch scoring accuracy problems? Or problems with bias?
After I scored the learners' responses to the questions, I copied their scores (Perfect answer = 2; Partially correct answer = 1; Wrong answer or blank = 0) into a spreadsheet. This process helped me work out the ease of answering for each question, i.e., which items the learners had acquired and which they had not.

Did you ask the students themselves to score their own test? Or a classmate's test?
Yes, I asked them to form pairs and score each other's test.

Do you think your test was reliable? What did you do to check?
I piloted the test with learners with similar proficiency and motivational levels and I analyzed the reliability of each item statistically.

Consider: Did you ask another teacher to look over your test before you administered it? What did he or she say? Did you make any changes to your test as a result?
As I explain above, I asked a colleague to check the items before I administered the test. My colleague commented that some of the words (not the target verb collocates) used in the sentences might be difficult for the learners and that the context of some of the sentences might not be familiar to all students. Based on this feedback, I replaced some difficult words with easier words, and I modified some of the sentences used in the test.

What did you do to ensure all learners had the same conditions under which they took the test?
I made every effort to create the same conditions for all students, including making an announcement and providing explanations about the test and administering the test in the same room in the same period for the same duration for all three groups of students. There were approximately 15 students in each of my classes.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Reporting scores

Did you report the scores to learners? If so, how did you report test scores to learners?
Yes. I accomplished this by asking pairs of learners to score each other's answers so that they could get a rough idea of their total scores.

What was your goal in reporting the test scores to learners?
One of my primary goals was to inform the test-takers how well or poorly they performed on the verb collocation test, i.e., what they knew and what they did not know. Also, I hoped that the results could raise their awareness of the importance of learning collocations while learning vocabulary.

Did you report the scores to anyone else?
I reported the test scores to my colleagues so that we could come up with effective ideas for learners to learn collocations effectively.

Did you spend time explaining scores or answering learners' questions about scores in or out of class?
As I explained earlier, after students took this test, I asked them to form pairs and exchange test sheets. Then, I elicited the answer to each question from the students and checked whether or not the answer was correct. Then I had them calculate the total score and return the test sheet to the test-taker. I collected the scored test sheets and scored the tests myself. Finally, I explained what the test scores meant in addition to mean, standard deviation, and percentile, and I answered the students' questions about the correct answers and how they could improve their knowledge of collocation.

Did you report peer-assessment scores or self-assessment scores on the test?
As test-takers assessed their peer's answers after taking the test, they could think back about their own likely score. Thus they indirectly assessed their own performance by looking at the scores they awarded their peers.

How quickly did you report scores to learners? Was speed a priority?
I asked all student-partners to score the answers by listening to my explanations so all test-takers knew the total score and what items they got correct or wrong immediately after they took the test. I believe that giving

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

feedback on performance as soon as possible is pedagogically important as learners should modify any imperfect knowledge as soon as possible.

Section Three: Using Test Scores

Cut scores

How did you decide which scores were passing or failing scores (cut scores)? How did you decide which scores meant a specific grade on a test?
I set the cut-off score at 70 following the Japanese testing convention used in higher education for intermediate- to upper-intermediate students. Actively knowing 70% of the collocations in question is good news for students, and suggests that they can use collocations fairly well, though there is still room for improvement. In general, students who get about 70% are usually happy and are motivated to improve their knowledge because it shows that they got a passing grade. In our university and probably in other universities too, if students get a score of 72, they can get a B grade (equal to or higher than 70), and if they get a score of 85, they can be given an A grade (equal to or higher than 80).

Did a language use framework such as CEFR or other standards help you determine cut scores?
While setting a cut scores at 70 is a common convention in Japanese higher education, I also considered CEFR descriptors as well as course materials and tasks used by the learners.

Did your institution stipulate cut scores?
Yes, my department sets cut-off scores depending on the type of test. Among these, the most important test is TOEFL ITP (Council on International Educational Exchange, 2019). The department imposes TOEFL ITP 480 as an exit criterion before students can advance to Year Three. My collocation knowledge test may help students improve their Listening and Structure and Written Expression scores on the TOEFL ITP. In other words, my verb collocation knowledge test could be labeled a "Pre-TOEFL Vocabulary and Grammar Test."

Did you consult a testing book or think of previous coursework you had to determine cut scores?
Yes. I used testing books, my colleague's suggestions, course materials, and tasks used by the learners to help suggest a cut score.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did a colleague or supervisor suggest cut scores?
Yes, my colleague suggested that 70% is an appropriate cut score for this type of a test.

Using test scores

How did you use learners' scores from this test?
I used the scores mainly for my teaching, but I shared the information with my colleagues so that we could get information about what verb collocations students were weak on. As long as my colleagues keep the test scores to themselves for reference or pedagogical purposes along with various strategies that prevent individual students from being identified, this information sharing is pedagogically useful and is often done in Japanese educational contexts. If the test scores were to be revealed to teachers outside the university, I am afraid that the test scores may have negative consequences for me, my students, and my university.

Did you hand the test back to learners? Did the learners get to keep the tests? Or did you take the tests back?
Yes, I handed the test back to learners and had them study the questions they got wrong, and then collected the test back from the learners.

Did you offer feedback to individual learners in addition to their test scores? Written? Orally? In or out of class?
I explained the answers and gave key explanations immediately after the test. I also answered questions from the learners.

Did you teach learners to interpret their scores?
Yes, I taught them how to interpret their scores by explaining the concepts of mean, standard deviation, and percentile.

Do you feel you learned what you needed to learn from the test results about what learners could and could not do, or what they knew or did not know?
Yes, I think so. I now have a rough estimate of their level of linguistic knowledge with respect to verb collocations.

Did learners ask you about the test itself (not the test scores) outside of class? If so, what did they want to talk to you about?
Yes, some of the learners did. They asked me how to improve their knowledge of verb collocations and how they can find out such

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

information. I suggested websites, effective listening and reading materials, and dedicated software.

Did learners' test scores change your teaching?
Yes, now I try to emphasize and summarize verb collocations at the end of each class if there is time.

Did you skip content because learners did well on your test?
No, because my classes are content-based, and I believe that even if the learners did well on the test, they may not have good ability with automatized verb collocates yet. Learners need to encounter the same verb collocates in different contexts to improve their fluency in reading, listening, speaking, and writing.

If you could turn back time, what would you change about your test? What would you change about your test administration?
Feedback from the focus group of learners showed that they thought the test was useful and fair. I think the test and its administration went well.

Did others (parents or administrators or colleagues) give you feedback on the test?
My colleague gave feedback on the test. I can now collect and "bank" additional collocates, including verb-noun, verb-adjective, and adjective-noun collocates as well as multiword units for future tests.

Section Four: Evaluating and Reviewing your Answers

To what extent do you think you've described recurrent patterns in your work with tests?
I have been developing tests that have not been much used in English language teaching in Japan in order to measure critical skills that should be paid attention to, including knowledge of multiword units and formulaic sequences. I believe these will improve learners' speaking and writing fluency. Such critical skills include automatized knowledge of verb collocates. Thus, the development of such tests is germane to my long-term research on the effects of tasks on the development of L2 oral fluency and interactional skills.

To what extent is your test here an innovation, or something new, for you?
I often notice learners' misuse of verbs in classroom tasks as well as speaking and writing tests, partly as a result of L1 transfer and partly

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

because of inadequate exposure and use of multiword units and collocates in EFL contexts. However, based on my teaching experience, while teaching such language items may raise learners' awareness of their importance, this has not been very effective in helping them acquire these language items, and I have never measured their knowledge of verb collocations to arrive at quantitative or qualitative data on learners' knowledge. Therefore, creating tests designed to measure learners' collocational knowledge and obtaining relevant data will help me understand which collocates they have learned and which they are weak on, thereby yielding new pedagogical insights for myself as well as for colleagues.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Providing an Oral Summary of a Written Text as a Mid-semester and Final Test

Meredith Stephens
Tokushima University
& Meagan Kaiser
Tokushima University

Introduction

Both of us are faculty members at a public university in Japan. Meredith is an applied linguist with degrees in applied linguistics, Japanese and linguistics, and education. She is interested in English language pedagogy in Japan, with a focus on how listening comprehension facilitates reading comprehension. Meagan is a TESOL specialist with degrees in TESOL and foreign language education with a specialty in Japanese education. Her main research interest is children's language pedagogy in Japan.

Tests in Japan often feature discrete language skills through exercises such as "fill in the blanks." This test requires learner production of extended discourse, in an attempt to prompt learners' inner speech in English. The skills focused on here are speaking, lexico-grammar, collocation, and extended discourse. The test we present here is a model for mid-semesterg and final exams, which themselves are based on weekly tests we give to learners.

Thus we present two tests:

1. A weekly vocabulary test which is graded, and follow-up oral summary tasks, which are not graded.

2. A mid-semester and/or final exam comprised of reading a text and giving an oral summary, without looking at it, very similar to the ungraded oral summary tasks done following the weekly vocabulary test.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

The learners were one class of second year Engineering students, and one class of second year Biology students, taking required Communicative English classes. Classes typically comprise twenty-five students, are held weekly for ninety minutes, and continue for a semester of sixteen weeks.

The two classes follow along to stories read aloud to them over the semester from a locally produced textbook *Mind the Culture Gap*, edited by Susan Balogh, of Shikoku University, and Jodi Lindsay (2017). This textbook consists of cross-cultural anecdotes produced by local Tokushima residents, and international visitors and residents of Tokushima. The anecdotes appear in English on the left of each page and the Japanese translation on the right.

The Tests

Weekly vocabulary test and oral summary activity. The weekly test prepares learners to take the mid-semester and final exams. Each week one of the stories in *Mind the Culture Gap* is read aloud to the class three times, as the students read the same story along silently as an assisted Repeated Reading exercise (Taguchi & Gorsuch, 2002). The learners are presented with a written text that has ten of the vocabulary items that have been substituted with synonyms. As students read along, they identify the original words and the synonyms which the teacher has provided. The teacher has changed ten words from the story to glosses or synonyms. The students write both the words in the text which have been substituted and the synonyms or glosses which the teachers have provided for each one. The teacher marks these tests weekly and these scores comprise 40% of the grades over the semester.

The objective was to cover both bottom-up and top-down skills, as Christine Nuttall explains in *Teaching Reading Skills in a Foreign Language* (1982). The vocabulary identification was a bottom-up skill, and the oral summary was a top-down skill. In Japan there is an imbalance in favour of bottom-up skills. Both are necessary, and bottom-up skills serve as a preparation for top-down skills. Therefore, first, students had to identify vocabulary in the weekly test by listening and reading (bottom-up, input), and then they had to produce an oral summary (top-down, output). A second purpose of vocabulary identification was to encourage focused, extensive, and repeated listening, as Meredith read the story to them three times while they identified her substitutions. Because they would be assessed, the students concentrated intensely as she read the story three times, and asked us questions after the readings to confirm their answers.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

See Table 4-14 for a sample weekly vocabulary test (text adapted from Herdman, 2017, p. 59).

Table 4-14: ***Sample weekly vocabulary test***

I have learned many things during my time in Japan. I've learned that Japanese people don't often say I love you, or give each other hugs. I've learned many different social rules that must be followed when people interact with one another. But one of the most interesting lessons that I've learned involves edamame.

My first time eating a Japanese meal at a restaurant, there were many bowls of edamame sitting on the table. Edamame are boiled, salted soybeans and are very popular at restaurants. I had never seen them before. All of my friends around me kept telling me they were delicious, but to me, they looked rather unappetizing, so I didn't eat any.

Finally one day when I went out by myself, I noticed that the waiter had brought some edamame to my table. Because there was no one around to judge my reaction, I decided to try one. Not knowing any better, I picked one up and simply bit into it - skin and all. Naturally, it was disgusting. I was convinced at that point that the reason I had never seen anyone eating them was because everyone else hated them, and had only been lying to convince me that they liked this terrible tasting food.

A few weeks later, I went out with my friend to another restaurant. The waiter brought a basket of edamame and I confronted my friend, saying that there was no way he could possibly like such a tough, salty and altogether gross food. He was very confused. He picked one up, opened the shell, and ate the bean inside. Needless to say, I was very surprised and felt quite foolish. I tried eating one the proper way, without the skin, and found that they are actually pretty good. I can still hear my friend laughing when he found out I had been eating the skins.

Highlighted words in text and matching synonyms
interesting (fascinating)
involves (concerns)
rather (fairly)
judge (test)
disgusting (awful)
hated (disliked)
gross (horrible)
needless to say (of course)
pretty (quite)
found out (discovered)

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

The students read the text but then hear the same text with alternate words, as in "But one of the most fascinating lessons that I've learned concerns edamame..." When the teacher is done reading the text aloud using the matching synonyms, the students hand in their sheet of paper with each of the original words from the text that has been substituted with a gloss or synonym (what they read), and their own substitutions (what they heard), as in: interesting/fascinating, involves/concerns, judge/test, etc.

After learners work with the vocabulary test and hand it in, students are then asked to engage in a lengthy oral summary activity of the text. First, the teacher highlights key lexico-grammar and collocations from the text for that day. The words and collocations selected for the next phase may include but are not limited to the substituted words in the weekly quiz. The teacher reads aloud the key collocations, tells the students where to find them according to the line and paragraph numbers, and has them highlight them themselves. Here is an example of one of the stories with the highlighted lexico-grammar and collocations (adapted from Herdman, 2017, p. 59). See Table 4-15.

Table 4-15: ***Sample oral summary text***

I have learned many things during my time in Japan. I've learned that Japanese people don't often say *I love you*, or give each other hugs. I've learned many different social rules that must be followed when people interact with one another. But one of the most interesting lessons that I've learned involves edamame.

My first time eating a Japanese meal at a restaurant, there were many bowls of edamame sitting on the table. Edamame are boiled, salted soybeans and are very popular at restaurants. I had never seen them before. All of my friends around me kept telling me they were delicious, but to me, they looked rather unappetizing, so I didn't eat any.

Finally, one day when I went out by myself, I noticed that the waiter had brought some edamame to my table. Because there was no one around to judge my reaction, I decided to try one. Not knowing any better, I picked one up and simply bit into it – skin and all. Naturally, it was disgusting. I was convinced at that point that the reason I had never seen anyone eating them was because everyone else hated them, and had only been lying to convince me that they liked this terrible tasting food.

A few weeks later, I went out with my friend to another restaurant. The waiter brought a basket of edamame and I confronted my friend, saying that there was no way he could possibly like such a tough, salty and altogether gross food. He was very confused. He picked one up, opened the shell, and ate the bean inside.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

> Needless to say, I was very surprised and felt quite foolish. I tried eating one the proper way, without the skin, and found that they are actually pretty good! I can still hear my friend laughing when he found out I had been eating the skins.

Students then study both the English story and the Japanese translation from the textbook that the except comes from (*Mind the Culture Gap*, Balogh & Lindsay, 2017).

After five minutes of silent study, Student A closes their book, and their partner, Student B, keeps it open. Student A summarizes the chapter paying attention to lexico-grammar and collocations. The students are told to retell the story in their own words, and not to learn it by rote. However they are encouraged to learn the collocations by rote, and to incorporate them into their retelling as much as possible. Student B listens to Student A, looking at the written text and providing prompts where necessary. Then A and B reverse roles. Next, the teacher organizes the students into groups of four by grouping them both vertically and horizontally. The teacher organizes the students into groups of four, consisting of pairs of students sitting next to each other and the pairs sitting behind or in front of them. See Table 4-16.

Table 4-16: ***How students are organized to take the weekly test***

Student A	Student B
Student C	Student D

Student A now summarizes the text to Student C, and Student B to Student D, and then they reverse roles. Finally Student A summarizes the text to Student D, Student B to Student C, and then they reverse roles.

Mid-semester/Final exam. The students are told that for the mid-semester test, they will have to provide an oral summary of one of the texts that has been covered in the preceding weeks to the teacher for a duration of five minutes, after a ten-minute preparation period. They will be shown the specific text that they will have to summarize ten minutes before the test. For convenience we re-use here the text, "The Edamame Experience," (Herdman, 2017, p. 59) which students would have studied in class. They should not memorize the text for the test, but rather summarize it using the same lexico-grammar, collocations and cohesive devices which should be recycled while giving the summary as much as possible. The teacher tells the students the test will be audio-recorded. The two teachers grade the tests individually, and then discuss learners' scores. Feedback is then provided to the students.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

This process is repeated over the second half of the semester with the remaining chapters of the book with a similar final test, but using a different text. See Table 4-17.

Table 4-17: ***Sample test text***

Sample test text	
Study the English text and the Japanese translation for ten minutes, paying attention to the key points in the story, the vocabulary, collocations, grammar and connectives. After ten minutes close the book, and be seated across the table from the examiner. Summarize the story in your own words, while incorporating the points above.	
The Edamame Experience by Benjamin Herdman	ぼくの経験　「えだまめ」 ベンジャミン　ハードマン
I have learned many things during my time in Japan. I've learned that Japanese people don't often say *I love you*, or give each other hugs. I've learned many different social rules that must be followed when people interact with one another. But one of the most interesting lessons that I've learned involves edamame. My first time eating a Japanese meal at a restaurant, there were many bowls of edamame sitting on the table. Edamame are boiled, salted soybeans and are very popular at restaurants. I had never seen them before. All of my friends around me kept telling me they were delicious, but to me, they looked rather unappetizing, so I didn't eat any. Finally, one day when I went out by myself, I noticed that the waiter had brought some edamame to my table. Because there was no one around to judge my reaction, I decided to try one. Not knowing any better, I picked one up and simply bit into it – skin and all. Naturally, it was disgusting. I was convinced at that point that the reason I had never seen anyone eating them was because everyone else hated them, and had only been lying to convince	日本にいる間、ぼくはいろんなことを学んだ。例えば、日本人は、やたらと相手にI love youと言ったり、抱き合ったりしないし、また、人々がふれあうときにはいろいろな社会的ルールに従わなければならない、ということも知った。しかし一番面白かったのは「えだまめ」の出来事だった。 日本で初めてレストランで食事をしたとき、テーブルの上に枝豆を入れたボールが並べられた。その枝豆は茹でてあり少し塩味がついていて、日本でのの季節にはどのレストランでも人気があるものらしかった。見るのは初めてだった。周りに座った友だちは「うまいぞ。食べてみな」とぼくに言ったが、おいしくなさそうだし食べる気がしなかった。 ある日、思い切ってぼくだけで

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

me that they liked this terrible tasting food. A few weeks later, I went out with my friend to another restaurant. The waiter brought a basket of edamame and I confronted my friend, saying that there was no way he could possibly like such a tough, salty and altogether gross food. He was very confused. He picked one up, opened the shell, and ate the bean inside. Needless to say, I was very surprised and felt quite foolish. I tried eating one the proper way, without the skin, and found that they are actually pretty good! I can still hear my friend laughing when he found out I had been eating the skins.	あのレストランに行った。あの時のウエイターがまたぼくの前に枝豆をいれたボールを運んできた。その時、ぼくの周りにはぼくの食べ方をとやかく干渉する奴もいなかったので、自分のやり方で食べてみようと枝豆に手を出した。そいつを皮ごと口の中に放り込んだ。しかしそれはひどくまずくて食べられたものではなかった。 数週間して、ある友だちとほかのレストランへ行くと例によって枝豆がやってきた。 ぼくがその友だちに「この塩味のするグロテスクな食物はまずくて食えないだろう」と言うと、その友人はひどく驚いた様子だったが、ひょいと一つ手に取ってさやをむいて、その中の豆を食べて見せた。ぼくが実にびっくりしたのは言うまでもない。彼がしたようにさやの中から豆を取り出して食べてみた。それで「ああ、ぼくってアホだよな～」としみじみ思い、いま初めて「枝豆って実にうまい物だな～」と悟ったのだった。ぼくには今でも「君はこのさやをとらずに、これまで食べていたの?」とお笑いしていた友だちの声が聞こえる。

Learners' audio-recorded tests are graded by two teachers individually, using the scoring criteria in Table 4-18.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 4-18: ***Scoring criteria for the mid-semester examination***

Assessment Criteria	Description
Inclusion of key points (5)	Is the student able to identify and include enough of the key elements to re-tell the story without changing the original meaning and without leaving the listener confused?
Collocations (5)	Has the student remembered the collocations accurately? Are the collocations used correctly in the story? Does the place where a collocation is used make sense in the story? Has the student correctly understood the meaning?
Lexico-grammar (5)	Is the student able to choose sentence structures and vocabulary that conveys the intended meaning?
Cohesive Devices (5)	Is the student able to make appropriate use of transitional phrases and repetition of key terms to bring the story together into a clear narrative?

Contributor's Questionnaire Responses

Section One: Test Planning and Writing

Why did you write the test? What were the purposes of the test? Japanese students lack the opportunity to produce extended discourse because speaking skills are rarely tested. This class was entitled *Communicative English* in English, and *Hasshingata Eigo* in Japanese. The latter refers to output, that is, speaking skills.

As I (Meredith) taught this class for over ten years I kept refining my methodology in order to foster speaking skills. I noticed that one of the major areas of difficulty was the collocations. Japanese collocations do not map neatly onto English collocations, but students tend to transfer the Japanese ones to their spoken English. I decided to have them produce spoken discourse which included collocations from the course book. I highlighted the collocations from the stories in the course book, and asked the students to memorize them and incorporate them into their extended discourse. I told them that memorizing collocations would help them in the future when they had to speak English, because they could recycle the collocations in their speech without having to assemble phrases using individual vocabulary. Memorizing collocations for later use would reduce learners' cognitive load.

Another purpose of the test was to lead learners to process English in its natural order. Japanese students tend to read and translate a text from right to left (Kato, 2006) in a process known as *kaeriyomi*. Much of

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

learners' "reading" of English tends to be reading and translating. Translation into Japanese involves changing the word order, which tends to be the inverse of English word order. The act of producing extended discourse leads students to engage with the natural word order of English in order to tell a story.

Consider: To help learners focus their thoughts, or change how they studied?
Yes. Please see above.

To award a course grade, or part of a grade?
Students undertook this test in the mid-semester (10% of their grade), and again at the end of the semester (10% of their grade).

To learn whether learners met a course objective?
The syllabus stated: "To develop listening and speaking skills through purposeful interaction."

To use the scores to give learners feedback on their progress?
The feedback from the mid-semester test was intended to help students practise collocations more effectively in preparation for the end of semester test.

Were you concerned at how long the test would take to administer?
Yes. The students were allotted three minutes to complete their extended discourse for both the mid-term and the final exams. Some students were fast and others paused frequently, and we were concerned about this variation in time. Several students went over the time limit, primarily due to multiple long pauses during the mid-semester test. At the end of the semester, a timer was used to enforce the limit and students were only allowed roughly 30 seconds beyond that to finish up their story.

Were you concerned at how long the test would take to score?
Yes. We scored the tests on the spot, and it was difficult to do this in the limited time.

Were you concerned how you might use the test items themselves for learner feedback?
Yes. Most students were nervous during the mid-semester and final exams. We didn't want to penalize them for their poor performance when the

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

reason was nervousness. We didn't want to give them negative feedback because this may impinge on their confidence and therefore performance.

Did you consider having your students take your test on a computer? Why or why not? If not, what were your concerns?
No. Speaking requires an audience. As an interlocutor speaks, they monitor the response of the listener's eye contact, facial expressions and posture. Speaking to a computer ignores these critical elements of everyday spoken communication.

Did you consider making the test an open book test?
No. We wanted the students to memorize the collocations and summarize the story as naturally as possible.

Did you plan to allow learners to re-take a test for improvement? The same test, or a different test?
Yes. Some of the learners performed poorly so we sent them back to practise with a friend and called them back to repeat the test at the end of the lesson.

What sources did you draw from for your test items?
Students had to provide an oral summary of selected stories from the course book. I adapted this idea from a writing textbook instructing on how to provide written summaries. Students had to read a paragraph, noting ten keywords. Then they had to look at the keywords to write the paragraph in their own words. I decided to apply this idea to an oral summary.

Test item ideas or content based on textbook activities?
The content of the test came from the coursebook. Students were asked to summarize and use collocations from stories they had read in the coursebook.

Test item ideas or content from course objectives?
The idea came from the desire to meet the goal of developing speaking skills through purposeful interaction.

Test item ideas or content from what learners do in class?
Yes. The students practised producing an oral summary of the text with different partners during the lessons. This practise was reflected in the test.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you consider learners' communicative competence when writing test items?
What aspects of communicative competence?
Yes. This test was designed to capture a student's spoken competence in the presence of two listeners.

Which test item formats do you prefer to use?
Performance test format.

What types of learner knowledge do you believe you are capturing in your test?
In principle, students must have a genuine understanding of the story as a whole and of the collocations they are being asked to use in order to create coherence in their summary.

What learner skills do you believe you are capturing in your test? How does that change with test item types you used on the test?
We are capturing skills used to make an oral summary, such as having an introduction and a conclusion, and use of collocations.

If you wrote a performance test (where learners have to converse or present a topic, or write something above the sentence level), how did you get the ideas for what task learners had to do for the test? Consider: From tasks learners do in class? From tasks they have do to in real life?
The Communicative English course is the last required English class for university students across all faculties at our school. They may need to use English in their various careers, be they in Engineering, Medicine, Science, Business, or Education. They may want to use English when they travel overseas. As Meredith was teaching this course she took an overseas trip during the break. Upon her return to Japan, the flight was suddenly cancelled and rescheduled. Passengers were redirected out of the airport to hotels. She witnessed the confusion of the many Japanese passengers as they were ushered out of the airport onto the different buses or given taxi vouchers, and at breakfast in the café the next morning. She realized how important real-time communication was and reaffirmed her resolve to keep promoting the skills required in this test.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

How did you get ideas on how to score learners' performances (the scoring criteria)?

After the mid-semester test, we refined our focus on our goals, and together decided on the following rubric: Introduction: 1, Collocations: 4, Pace: 2, Coherence: 2, and Conclusion: 1. As below in Table 4-19.

Table 4-19: ***Revised scoring criteria for final exam***

Assessment Criteria	Description
Introduction (1)	Is the student able to introduce the characters and set up the frame of the story for the listener?
Collocations (4)	Has the student remembered the collocations accurately? Are the collocations used correctly in the story? (Does the place where a collocation is used make sense in the story? Has the student correctly understood the meaning?)
Pace (2)	Is the student able to tell the story fluently without an overabundance of hesitation or rush?
Coherence (2)	Is the student able to choose sentence structures and vocabulary that conveys the intended meaning? Is the student able to use cohesive devices to make the story as a whole easy to understand?
Conclusion (1)	Is the student able to finish the story and bring it to a natural sounding close?

If you used points on a scale for scoring learners' performances, how did you decide on how many points on a scale to use?

The test was worth 10% of the semester grade. The point total added up to 10 (1 + 4 + 2 + 2 + 1 = 10) and the sections of the rubric were weighted by relative importance to communicative competence. Collocations were weighted most heavily, followed by pace and coherence, and finally the introduction and conclusion were given only one point each.

We realized after the midterm that we would likely have a more effective measure of the communicative competence we were trying to assess if we weighted the criteria. Also, setting up the rubric in a way that more closely mirrored the arc of the story seemed like a clearer and more effective way to explain to students what we hoped for them to accomplish.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Were you concerned about whether you could get a colleague to help you score learners' task performances?
Yes. We scored the tests independently. The final score was the average of the two.

Did you change your plans or your test to reflect time constraints in terms of test scoring?
Yes. In the end-of-semester test we measured and enforced the three-minute rule more effectively by using a timer.

Did you plan to give the scoring criteria to the learners for future self- or peer-assessment?
We gave the students the scoring criteria two weeks before the test before both the mid-semester and final exams.

Did you seek help from a peer to clarify what your test items or tasks were measuring?
Yes. The criteria emerged as a result of our discussions.

Did you make any changes to your test or items as a result of your colleague's feedback?

Yes. We decided to give more weight to the collocations, because this was the most important feature of our test.

Did you compare your test to the lessons that learners had?
Yes. The students practised producing an oral summary of the text with different partners during the lessons.

Did you compare your test to the textbook or other materials learners used?
Yes. We created the test from the stories contained in the course book.

Section Two: Test Administering and Scoring

Were you concerned about test security? What did you do to ensure test security?
Students were told two weeks before the test that they would have to summarize one of three selected stories. However, they were not told which story they had to summarize until they entered the classroom.

We wrote the three chapters containing the stories that the students had to summarize on the board before the test. The first third of the students

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

had to summarize the first story during the first thirty minutes of the ninety-minute lesson. The second third of the students had to summarize the second story in the second thirty minutes of the lesson, and the last third of the students had to summarize the third story in the final third of the lesson. Therefore, though not a security problem, it was an issue that the students who were tested in the beginning of the thirty- minute segment had less time to prepare for their specific test than the others.

How did you deal with learners who missed the test, or who were late for the test?
They were told to come and take the test again the following week.

How did you prepare learners to take the test? Were there any test items or test procedures that were new to them?
Students practised making oral summaries of stories to various partners in the lessons before the test.

Did you pilot your test? Do a trial run? Did the pilot result in any changes to the final version of your test?
No, but we did adapt the scoring criteria based on the results of the mid-semester test. The assessment criteria of the pilot were vocabulary and a comprehensible beginning, middle and end of the story, which were evenly weighted. We changed the final version of the test scoring rubric to evaluate collocation rather than vocabulary, to give more weight to collocation than the other criteria, and to include pace and coherence.

As mentioned earlier, setting up the rubric in a way that more closely mirrored the arc of the story we hoped would help students to better understand what we were asking them to accomplish. It was made explicit that students need a proper introduction, must use the collocations accurately and often in the telling of their story, and the students must have a conclusion. We realized after the midterm that we would likely have a more effective measure of the communicative competence we were trying to assess if we weighted the criteria. Collocations were weighted more heavily than other criteria in order to make clear their importance to the text and to get a more accurate assessment of a student's skill in using the collocations specifically. Pace was included to give a way to assess fluency in performance that we were missing previously. Lexico-grammar and cohesive devices merged into the category of coherence for the final test.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you write any of the test in the learners' first language? Why?
The course book consisted of stories in English and their Japanese translations. The students were instructed to read the stories in Japanese before they read them in English, in order for them to get an overview of the content. The first language was used to provide background knowledge, which we thought would facilitate top-down processing skills. We wanted them to consider the story as extended discourse.

Was your test administered on a computer?
No. We considered embodied communication to be too important to do this on a computer. Also, these days there is an abundance of computer learning and many learners find this to be a chore.

For classroom tests: How did you accomplish scoring learners' tests? Did you write a test key?
The weekly vocabulary tests were a stepping-stone to the later requirement to provide an oral summary. They provided a chance to both focus on vocabulary and to listen to extended discourse. The vocabulary and the listening provided the groundwork for their oral summaries. Students were free to choose which collocations to memorize.

Did you consider alternate answers and add them to the key?
Students did not need to strictly conform to the exact collocations in the stories.

Did you randomize learners' tests for scoring, thus erasing any order from where learners sat, or at what point they handed their test in?
The three-minute mid-semester and final tests were done in the order of the students' names on the roll.

For performance tests: How did you accomplish scoring learners' performances?
We scored them separately as a team. We did not compare scores until after the test session. We did not record them, because this would have added to the students' nervousness.

We scored them at the same time as the performance or in the brief interval as one student finished and the next started. During the mid-semester test we experienced fatigue because we could not manage our time efficiently. During the end-of-semester test we did not experience so much fatigue because we managed our time more effectively.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Do you think your test was reliable? What did you do to check?
Although we scored independently, the scores resulting from our rubrics were very similar, so we think it was reasonably reliable. We compared scores after having scored the tests independently and discussed any cases where we were on the fence about the score.

Reporting scores

Did you report the scores to learners? If so, how did you report test scores to learners?
We reported the scores to students individually the week after the test.

What was your goal in reporting the test scores to learners?
Prompt feedback is essential to learning.

Did you teach learners how to interpret their test scores?
The students were given the assessment criteria before the test.

Did you report the scores to anyone else?
No.

Did you spend time explaining scores, or answering learners' questions about scores in or out of class?
We sent the students an email outlining the assessment criteria before the test.

How quickly did you report scores to learners? Was speed a priority?
Speed is a priority but we only see students weekly so we had to wait until the next week to be able to give feedback in person.

Section Three: Using Test Scores

Cut scores

How did you decide which scores were passing or failing scores (cut scores)? How did you decide which scores meant a specific grade on a test?
The passing grade at the university is 60%. Our mid-term and end-of-semester tests comprised 10% each. These grades were combined with other assessment criteria (80%) to determine the overall grade.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did learners' test scores have any positive or negative consequences for you, in terms of your institution?
Scores were for the students' and for our use only. The test scores have no consequences for us in terms of our institution beyond being a portion of the score considered for the overall course grade.

What was the role of the test score in determining learners' grades?
The combined test scores constituted 20% of the students' grades.

Were other measures used to decide learners' grades, besides your test?
40% of their grade consisted of weekly listening tests to live English. The teacher read a story aloud every week and substituted ten words for their synonyms. The students had to identify the synonyms.

Another 40% of their grade consisted of an online extensive reading and listening program, Xreading® (2019), a commercial online library that gives students unlimited access to over 1000 graded readers from international publishers, combined with a learner management system that allows teachers to track their students' reading and listening progress (Paul Goldman, March, 2017, personal communication). Students had to read and listen to stories totaling 50,000 words each semester and achieve a grade of at least 60% on comprehension tests.

What was the relationship of the other measures to your test?
Both of these measures involved extensive reading and listening, the first to live English and the second online. The content of the first measure, the weekly listening test, was the same as the stories the students were required to provide an oral summary of in the test.

The content of Xreading® was not prescribed. It did not directly prepare students for the test, but rather helped habituate them to extensive reading and listening to English.

Did your test capture some knowledge, skill, or ability the other measures did not capture?
The weekly classroom activities required comprehension skills, but the test captured productive skills.

Reporting scores

How did you report scores to learners? Was timeliness of concern to you?
We reported the scores to the learners the week after the test. Timeliness was of concern, but this was the soonest opportunity I had.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you offer feedback to individual learners in addition to their test scores? Written? Orally? In or out of class?
Feedback was oral and in class. Japanese Engineering students tend to be reticent and they tended not to overtly respond to my feedback.

For performance tests, did you use the test criteria to help learners interpret their scores?
Yes, that would be an accurate description.

Using test scores

Do you feel you learned what you needed to learn from the test results about what learners could and could not do, or what they knew or did not know?
We learned how important it was to give students an individual opportunity to talk in English without interruption. We learned how big individual variation in test preparation and performance was. We were surprised that some students had skills that were only evident when they were given the opportunity to speak individually.

Did your test change how learners studied? Consider: Did you use learners' scores to find out if your test caused washback?
We instructed the students in how to prepare for their oral summary, not by just silently reading the story, but by providing an oral summary to a partner who could see the story in the course book in both English and Japanese. We think this practice caused washback within the participating community of students.

Did you use particular item types or a performance test to change learners' practices or support their learning? Did their scores indicate they had changed their learning practices?
We changed the focus from vocabulary for the mid-semester test to collocation for the final test, and gave an increased weighting to collocation. The final test revealed the students had paid increased attention to collocation.

Did learners' test scores change your teaching?
It made us realise how important giving students a chance to speak was, rather than dominating the classroom with our own talk. It also made us spend more time on stressing the need to memorize collocations, and to explain to them why it was important. We told students in their first

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

language that a wide vocabulary was necessary, and that knowledge of collocations would make it easier to communicate quickly. We told them that this was a useful skill for their future rather than just being useful to pass the test.

Did you re-teach content because learners didn't do well on your test?
We repeatedly taught students how to choose the key words from the story to make a summary, and the importance of choosing collocations to remember rather than directly translate those ideas from Japanese.

Did you change your teaching for future courses based on test results?
We have been convinced of the importance of memorizing collocations for many years, rather than translating with the Japanese-English dictionary. This has been affirmed by this test, as we observed certain students mastering the collocations in their oral summaries.

Section Four: Evaluating and Reviewing your Answers

To what extent do you think you've described recurrent patterns in your work with tests?
Meredith has been concerned with how to teach collocations for over fifteen years. We told students to incorporate collocations in previous performance tests, such as those of paired conversations. This most recent test is unlike our previous tests, because of the increased weighting given to collocations. Previous conversation tests required more creativity than the current test. The current test was not about producing an original text, but rather about summarizing an existing text.

To what extent is your test here an innovation, or something new, for you?
For Meredith, this test is a result of innovations over more than fifteen years. Every year she evaluates her teaching, and uses feedback from students about their learning. Over the years she has had ample time to generalize about the areas that need most improvement. This test has been tailored to remedy these areas of weakness.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

An Oral Voicethread Test for First-semester French Language Learners in a U.S. University

Beatriz García Glick
The Pennsylvania State University, Hazleton

Introduction

I am an Associate Teaching Professor at the Pennsylvania State University, Hazleton campus, where I have been teaching for the past nine years. I teach both French and Spanish at the Beginning and Intermediate levels. My research interests include the use technology to enhance language teaching in terms of improving students' motivation as well as for assessing their progress. A recent article I published is "Improvement of Present Subjunctive Oral Production in Graded VoiceThread Tasks," (2016), where I explored "VoiceThreads" (VoiceThread, 2019) to improve students' oral production and understanding of the Present Subjunctive in Spanish.

I first became interested in VoiceThreads in 2014 because it was supported by my University and because it is a web-based application that allows students to record a conversation, to listen to themselves, to each other, and to the instructor's feedback. VoiceThread also allows images and text to be added so it allows each student to make their contribution as creative as possible. In many instances, students include a picture of their best friend so we can compare their oral description with the image on the screen. I typically use VoiceThread at least once a month and have demonstrated its usefulness in improving learners' production of the Present Subjunctive in Spanish.

I like to assess my students for achievement using a variety of technology tools including blogs, VoiceThread recordings, online discussions, chats, and community projects where we use a recording studio to upload a short lesson/ skit/ to a public blog for the community. I provide objectives, rubrics, feedback, and grades on all of the assessments.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

The Test

VoiceThread Activity

Objectives: 1.) To create and write a short presentation in the Present Tense about your daily routine and your plans for the Fall; 2.) To practice speaking; 3.) To listen to your classmates and to learn from them; and 4.) To receive feedback on written and pronunciation skills.

Level: Beginning French

Activity: To assess Speaking and Writing Skills by producing a recording using the Cloud application "VoiceThread" (https://voicethread.com/).

VoiceThread Title: *Ma routine et l'automne prochain.*

Due: _____ at midnight. You can make revisions to improve your grade.

You have learned in Chapter 1 the Present Tense of verbs in *er* and in Chapter 2 the immediate future using *aller*. Let's discuss in two paragraphs: 1.) What do you do? What is your normal routine? What is the routine of your parents? and 2.) In the Fall, what are you going to do ?

Six sentences minimum per paragraph. Include specific details, like the time that you eat lunch.

For the first paragraph: *Qu'est-ce que tu fais normalement, le lundi/ le mardi/ / le week-end?* Choose one time frame.
Use the Present Tense of: *déjeuner* (to have lunch) / *dîner* (to have dinner) / *ne jouer pas* (to not play) / *ne téléphoner pas à* (to not telephone)

Par exemple,
Le lundi matin (Monday morning), *je* . . . (I)
Le lundi après-midi (Monday afternoon), *je* . . . (I)
Le lundi soir (Monday evening) *je* . . . (I)

Que font tes parents normalement le lundi, le mardi/ le mercredi/ le week-end? What do your parents do normally on Monday, on Tuesday, on the weekend?
Le lundi matin (Monday morning), *ils* . . . (they)
Le lundi après-midi, ils . . . (they)

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Le lundi soir, ils . . . (they)

For the second paragraph: *Qu'est-ce que tu vas faire l'automne prochain?* What are you going to do next Fall?
Structure: Present tense of « *aller* » + infinitive, the Immediate Future
Answer all questions and ask one question to your classmates.

Vas-tu étudier les sciences physiques comme l'astronomie? Are you going to study the Physical Sciences like Astronomy?
Vas-tu étudier l'informatique? Are you going to study Computer Science?
Vas-tu étudier les sciences naturelles comme la biologie? Are you going to study the Natural Sciences like Biology?
Combien de cours est-ce que tu vas suivre (follow)?
Conclusion: *C'est tout.* (That's all.). *Merci pour votre attention.*

Note linguistique:
In English, we take courses; In French, we follow *(suivre)* courses.
In English, we take an exam; In French, we pass (*passer*) an exam.
In English, we pass an exam; In French, we succeed (*réussir*) an exam.

Upload your document and record your voice. *Bon courage!*

Rubric

VoiceThread assignments are online records of your progress in speaking skills during the course. You will complete two VoiceThread assignments. Each VoiceThread assignment is 10% of your final grade and needs to completed by due date on your Course Outline at the end of the syllabus. You may view these topics by opening the VoiceThread program and opening the French course group. You will prepare a slide (7 – 10 sentences), save it on the Desktop, and upload it to VoiceThread by clicking the upper left hand corner icon with horizontal bars on the reference slide. Click "Edit," and "Add Media"). Then, select the uploaded slide, and click "Comment" to record on the microphone icon. To receive full credit, you need to upload a written slide and record your voice. The main purpose of these assignments is to improve your French writing and speaking skills. Comments will be added to your VoiceThreads by the Instructor. No late VoiceThreads will receive credit. Maximum of five points. You will receive feedback on your writing and recording so that you can submit a second recording for full credit.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 4-20: *Scoring rubric given to students*

0	1	2	3	4	5
No written information in French.	Slide with 1 - 2 sentences in French.	Slide with 3 4 sentences in French.	Slide with 5 6 sentences in French.	Slide with 7 or more sentences in French.	Slide with more than 7 sentences in French. No misspelled words.
Non-comprehensible voice Recording. **Past due date.**	Recording with 1 2 comprehensible sentences in French.	Recording with 3 4 comprehensible sentences in French.	Recording with 5 6 comprehensible sentences in French.	Recording with 7 or more comprehensible sentences in French.	A full one minute of comprehensible spoken French.

Contributor's Questionnaire Responses

Section One: Test Planning and Writing

Why did you write the test? What were the purposes of the test?
The objectives of the were: 1.) For students to create and write a short presentation in the Present Tense about students' daily routines and their plans for the Fall; 2. For students to practice speaking; 3.) For students to listen to their classmates and to learn from them; and 4.) For students to receive feedback on their written and pronunciation skills.

How did you decide how many subtests to write? Was there some correspondence between a subtest and a course objective, or was there some other reason to create the subtests you did?
The test had two subtests because in one paragraph or subtest, students had to speak in the Present Tense, and in the second subtest or paragraph, the students had to speak in the Immediate Future. Practice in both of these tenses were course objectives.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you write as many items or test tasks as you needed, or did you write more than you needed, then pare the number down?
In general, I always write more than I need. Then, I pare the number down to make the task more accessible in terms of time available as well as the vocabulary and grammar structures needed.

How did you decide how many items to write for each subtest?
Usually I ask students to give me several examples on a topic to make sure that they understand the concepts, and that they use the necessary vocabulary. Also, I ask that they try to speak in paragraphs and use various forms of the Present Tense such as "I" and "they."

How did you decide how many test items to write in total?
In general, enough for a paragraph. So for each task at a Beginner level, students have to complete three items pertaining to each task.

Did you have one version of your test, or did you create a second equivalent version?
Usually, I have several versions. In this test, students are asked to explain what they will do in the Fall because it is a Summer course, but if I teach the course in the Fall, I ask the students to explain what they will do over the Holiday Break, and so forth.

Were you concerned at how long the test would take to administer?
No, because I know how long it takes in general to speak about ten sentences in a beginning language course. Also, I know that students practice their assignment before recording the final version. This is rehearsed speech.

Were you concerned at how long the test would take to score?
No, because I have a rubric with clear parameters that evaluate the comprehension, duration, and accuracy of the students' rescordings.

Were you concerned how you might use the test items themselves for learner feedback?
I try to give personal feedback so that each student can improve their pronunciation skills. VoiceThread, a cloud application, allows me to listen to the student and to give her/him personal feedback. I use VoiceThread because my university has an institutional license, so there are no costs to my students or to myself.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you consider having your students take your test on a computer? Why or why not? If not, what were your concerns?
VoiceThread is a Cloud application that allows students to upload a slide and record speech to the slide. I find it very useful because it helps me to address each student's needs individually.

Did you consider making the test an open book test?
This test is open book. Students may use any reference material that they need.

Did you consider allowing learners to use additional sources such as dictionaries, or their notes, while taking the test?
Yes, when making their threads, students have access to textbooks, notes, and dictionaries.

Did you plan to allow learners to re-take a test for improvement?
Yes, students are asked to make improvements on their threads as part of their Oral Exam which is taken at the end of the semester and which incorporates speech from presentations made in VoiceThreads.

What sources did you draw from for your test items?
Consider: Test items or content from the ACTFL (American Council on the Teaching of Foreign Languages) guidelines or descriptors?
I use ACTFL guidelines and descriptors when testing oral competence in my students.

Which test item formats do you prefer to use?
In general, I like to give various types of tests including short answer, multiple choice, and performance tests such as the one in VoiceThread.

What types of learner knowledge do you believe you are capturing in your test?
I am addressing visual and auditory learners.

What learner skills do you believe you are capturing in your test?
I am addressing creative thinking, communicating, and collaborating.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

If you wrote a performance test (where learners have to converse or present a topic, or write something above the sentence level), how did you get the ideas for what task learners had to do for the test?
Yes, I try to implement tasks that are relevant and practical in everyday life. I also follow the ACTFL guidelines.

How did you get ideas on how to score learners' performances (the scoring criteria)?
The scoring criteria is based on a 0 – 5 point system which I have seen in various language textbooks.

If you used points on a scale for scoring learners' performances, how did you decide on how many points on a scale to use?
I find that for a quick assessment, a scale of 0 – 5 points works well because it follows a simple set of directions that allows students to understand the minimal expectations of the task.

Were you concerned about whether you could get a colleague to help you score learners' task performances?
I did not consider asking a colleague to help me evaluate the oral production of my students because I don't have any colleagues at the campus that are proficient in the language of the task.

Did you plan to give the scoring criteria to the learners for future self- or peer-assessment?
Students have access to the rubric since it is posted in the syllabus. When I give them feedback, it is directly related to the task as well as to the rubric requirements.

Did you compare your test to the lessons that learners had?
Yes, I made sure that the test specifically addressed the concepts that we were learning.

Did you compare your test to the textbook or other materials learners used?
Yes, I made sure that the test was relevant to the textbook and material covered in class.

Adapting existing tests
Are you required to use specific tests in your program?
No, at the branch campuses in my institution, we can choose our textbooks ourselves. However, since students often move to the main campus, I try to

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

use the textbook that is used there to make the transition easier for my students.

Section Two: Test Administering and Scoring

Were your tests photocopied, or did learners see your test items or test tasks another way, such as on the blackboard?
Learners visited the website called "VoiceThread" to view the requirements of this test.

How did you deal with learners who missed the test, or who were late for the test?
In general, I give students a one-week window of opportunity to hand in the test after the due date. This is explained in the syllabus of the course.

How did you prepare learners to take the test? Were there any test items or test procedures that were new to them?
Students worked in class assignments that were similar to those required in the test.

Did you write any of the test in the learners' first language? Why?
Yes, I wrote the instructions in English and gave examples in the target language. I wrote the instructions in English because I wanted the assignment to be clear.

Was your test administered on a computer?
The test was performed on a computer. Before submitting the recording for a grade, I met with my class in a Technology Laboratory and we reviewed the goals of the test, the ways of uploading the information, and began a practice test in class to show them how to upload the information. Although we reviewed the steps in class, there were still one or two students who did not share the recording with the rest of the group, and who did not realize this until they earned a "0" for the test.

For performance tests: How did you accomplish scoring learners' performances?
In this VoiceThread test, it is almost impossible to score them anonymously because their names appear on the recording as soon as you open it. There is also a written slide with their name on it and the script of their monologue. In order to make a test reliable, I assign specific parameters such as in the first paragraph, include the Present tense of certain verbs like *déjeuner* (to

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

have lunch), *dîner* (to have dinner), *ne jouer pas* (to not play), etc., as well as set a minimum number of sentences which are needed for full credit. This way, by being very specific, I try to eliminate bias in my grading and to focus on the pronunciation of specific verbs and constructions so that everyone is graded on the production of similar speech.

I then refer to the comprehensibility of their speech based on the quantity of intelligible sentences as stated in the Rubric. For example, three to four comprehensible sentences are worth two out of five points. A full minute of speech, which includes all of the parameters of the task, is worth full credit.

Because research shows that students learn more if given a second chance to review and submit the test, I also offer the students the choice of posting a second recording after I have given them feedback to earn full credit for this test.

Since the VoiceThreads are recorded speech, I review them at least twice to make sure that I give the students proper credit for all of the required material. I also review the written slide or script that accompanies their speech so I can give them advice on any misspelled words, faulty syntax, or other incorrect expressions.

There is no colleague or student aide who has knowledge in French who could help me score these tests. I have used Spanish/ English bilingual student aides to help me rate the comprehensibility of speech in VoiceThreads in Spanish classes that I teach, but I don't have any colleagues who speak French.

Do you think your test was reliable? What did you do to check?

The reliability of this test rests on the fact that all of the students have to address the pronunciation of specific verbs, constructions, and vocabulary. By making certain structures mandatory, all students begin by recording similar sentences but those who are more creative have the option to expand and make more complicated structures by adding adjectives, adverbs, and prepositional phrases.

This test is reliable because it holds all students accountable to produce a minimum number of structures by creating sentences with specific verbs and by answering specific questions.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Reporting scores

Did you report the scores to learners? If so, how did you report test scores to learners?
Yes, students were able to see their scores on the school's course management system. Students were also able to hear my comments about their pronunciation on the VoiceThread program. The comments were made in both English and French. I use English to explain what could be improved and French to model how it can be improved.

What was your goal in reporting the test scores to learners?
I wanted students to improve their pronunciation of French.

Did you teach learners how to interpret their test scores?
I referred students to the rubric.

Did you report the scores to anyone else?
No, only to students.

Did you spend time explaining scores, or answering learners' questions about scores in or out of class?
No, the rubric was self-explanatory.

How quickly did you report scores to learners? Was speed a priority?
Yes, speed was a priority because we only teach for 15 weeks and students have many concepts to understand. I gave them feedback one week after the submission of the test.

Section Three: Using Test Scores

Cut scores

How did you decide which scores were passing or failing scores (cut scores)? How did you decide which scores meant a specific grade on a test?
I determined the passing grade as 60% of the number of sentences that they had to produce because 60% is the passing grade for the course.

What was the role of the test score in determining learners' grades?
In general, a VoiceThread recording is worth 10% of the student's final average. This test is a measure of a student's ability to pronounce and express themselves correctly using some verbs in the Present Tense.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

How did you report scores to learners? Was timeliness of concern to you?
Yes, timeliness was a concern because I gave students the opportunity to resubmit the test for extra points.

Did you hand the test back to learners? Did the learners get to keep the tests?
The recordings and feedback were open to students during the entire semester.

Did you offer feedback to individual learners in addition to their test scores?
I offered oral feedback individually and then I reviewed orally in class the major mistakes that were made by most students.

For performance tests, did you use the test criteria to help learners interpret their scores?
Students knew that the objectives were to speak in the Present Tense about their daily routine and that of their parents and that they had to explain what they were going to do in the Fall. There was a minimum number of sentences expected as well as the use of certain expressions. I find that, by setting clear objectives and a minimum required amount of speech, most students understand the expectations and reasons for the task.

Using test scores

Do you feel you learned what you needed to learn from the test results about what learners could and could not do, or what they knew or did not know?
Yes, it is usually easy to grade this type of task because the objectives are clear to the students.

Did your test change how learners studied?
A majority of students resubmits their recordings for a higher grade which indicates that they value my feedback and that they try to improve their pronunciation and sentence structure in French.

Did you spend time going over the test in class?
Yes, in class, I addressed common mistakes and specific concerns of the students.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did learners ask you about the test itself (not the test scores) outside of class?
No, in general, students do not have specific issues on this test.

Did learners' test scores change your teaching?
Yes, after I reflect on their answers, I do include more tasks that address these issues if a majority of the students show a lack of understanding on how to produce the Present Tense.

If you could turn back time, what would you change about your test?
In general, students enjoy recording on VoiceThread because it lowers their anxiety level in speaking French. Students choose which recording to save so they can practice on their own before submitting a final version of rehearsed speech. On the other hand, it is not authentic oral production but at a beginning level, when most of the language learning relies on memorization, the use of VoiceThread activities encourages students to speak and to hear their classmates as well because the threads are open to all. This is also a motivating factor when they hear their classmates and compare them to their own production.

Section Four: Evaluating and Reviewing your Answers

To what extent do you think you've described recurrent patterns in your work with tests?
In order to teach grammar using authentic contexts, I follow the method described by Donato and Adair-Hauck (2002) called the PACE model: Presentation/ Attention/ Co-construction/ and Extension. This test is part of the "Extension" process of understanding grammar structures in authentic tasks. It asks the students to describe their routine using the Present Tense of specific verbs but students can add their own sentences as needed. I test student understanding of the formation of the Present Tense by asking them to also describe what their parents do during the week.

I prepare these tests after students have been presented with target structures in authentic texts such as radio conversations, television announcements, songs, paragraphs from a novel, which students have to complete using a particular structure such as the Present Tense. Then, we review their answers together in class and I prepare a test that extends their knowledge by asking them to practice the same structures that they have learned, but in a different task.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

To what extent is your test here an innovation, or something new, for you?
I have been using Voicethread recordings as performance tests for my students for the past four to five years. I have always found them useful in my teaching because they lower the anxiety level of my students, and because they are permanent recordings which can be graded at the instructor's convenience. The fact that students can listen to each other and comment on each other's thread are also positive features.

In the current test, an innovation for me is the fact that I recently began to give my students a second chance to record their presentation for extra points. In my opinion, the students benefit greatly from this repetition because they pay attention to their pronunciation and to sentence structure as well as to the communicative value of their task.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

A Speaking Fluency Test for Intermediate-Level German Using a Rubric Based on Grice's Conversational Maxims

Annis N. Shaver
Cedarville University

Introduction

I currently teach all of the German courses offered at our small private Midwestern university. We offer a minor in German which includes two semesters each of elementary-, intermediate-, and advanced-level language study. A search for reasonable tests of speaking fluency for intermediate level students, and particularly rubrics, or scoring criteria, for such tests, led me to create a rubric based on Grice's Conversational Maxims (Grice, 1975).

Our intermediate level of German is taught in two semesters, the first being a four hour per week course taught in the Fall semester and the second being a three hour per week course taught in the Spring. I use the textbook *Kaleidoskop* (Moeller, Berger, Wieden, Mabee, & Adolph, 2017), and follow the textbook's themes. I test ten times during the two semesters at theme/chapter intervals appearing in the textbook (six in Fall; four in Spring). Students are encouraged to speak German at all times in the classroom, and they achieve this goal approximately 90% of the time. Evaluating their speech on a continual basis is, however, unwieldy, and in some regards unfair, as students grapple with their shifting Interlanguage. Thus, in an attempt to fairly evaluate students' spoken language proficiency, I set aside one class meeting during each unit for targeted group discussion of the thematic material contained in the unit. The textbook provides ample discussion topics and situations, which I modify to fit the interests of the students and to allow them to address different perspectives.

Testing for speaking proficiency meets course objectives. All intermediate language courses at our institution use the same objectives

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

regarding language proficiency: "Apply basic German language skills appropriately through written and spoken communication," and "Synthesize communication skills appropriately within the cultural context of German-speaking peoples."

Evaluating spoken language proficiency in language learners is difficult and often overly subjective because of the influence of factors such as pronunciation, grammar, and manner, or affect, that influence the evaluation. As part of awarding a grade for spoken performance in the intermediate level of German, it was necessary to develop a rubric so as to moderate these superficial influences in order to evaluate true linguistic proficiency, that is, the ability to speak to a topic, and to participate in conversation about a topic of interest to all concerned. Thus, the teacher/evaluator must have a rubric that will ameliorate the tendency to allow pronunciation, grammar and affect to overshadow the student's demonstration of oral language proficiency.

The Test

I designed a rubric based on Grice's four Conversational Maxims to assess a student's ability to speak to a pre-assigned prompt, in an impromptu manner. I should note that the intermediate German classes are usually small (+/- 10), so this assessment is manageable. Also, students are assigned the prompt before the date set for the discussion, and are encouraged to practice with other students in the course. Responses should not appear to be memorized and only notes that contain vocabulary words are allowed during the discussion. Grice's four Conversational Maxims were instantiated into my rubric (Grice, 1975). The Maxims are as follows:

- The Maxim of Quantity is assessed through the number of turns taken in the whole class discussion. Each student is required to respond to the discussion prompt, to pose a question related to the prompt, and to respond to most (at least half) of the questions posed by other students.
- The Maxim of Relevance is assessed through the relevance of the student's responses to the topic of discussion. Responses should not be off topic without clear or sufficient connection to the discussion.
- The Maxim of Quality (modified from a focus on truth) is assessed through the student's grammar and pronunciation. Comprehensibility plus clarity, as well as a consideration for grammatical points covered in class, are taken into consideration with more concession

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

given for non-standard usages in the early part of the first semester than for such usages in the later part of the second semester.

- The Maxim of Manner is assessed through the clarity of the student's responses, including appropriate word choices, lack of ambiguity, appropriate register, and consideration for the audience.

Having found it difficult in the past to accurately assess conversation or discussion, I have found that this rubric provides an accurate assessment that allows me to determine which aspects of language, besides grammar and pronunciation, need more instructional or individual focus. Additionally, assessing through group discussion instead of individual speaking tests, causes students to negotiate the comprehensibility of their own speech production in a manner similar to that of real life. See Table 4-21 for the rubric.

Table 4-21: ***Rubric for discussions***

	5	4	3	2	1
Student contributes enough information to the discussion The Maxim of Quantity is assessed through the number of turns taken in the whole class discussion. Each student is required to respond to the discussion prompt, to pose a question related to the prompt, and to respond to most (at least half) of the questions posed by other students.					
Comments:					
Student's contribution is relevant to the discussion The Maxim of Relevance is assessed through the relevance of the student's responses to the topic of discussion. Responses should not be "off topic" without clear or sufficient connection to the discussion.					
Comments:					

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Student contributes clearly: pronunciation, grammar The Maxim of Quality (modified from a focus on truth) is assessed through the student's grammar and pronunciation. Comprehensibility plus clarity, as well as a consideration for grammatical points covered in class, are taken into consideration with more concession given for non-standard usages in the early part of the first semester than for such usages in the later part of the second semester.					
Comments:					
Student's contribution clearly answers or addresses the question The Maxim of Manner is assessed through the clarity of the student's responses, including appropriate word choices, lack of ambiguity, appropriate register, and consideration for the audience.					
Comments:					
Total Score:	/ 20				

Discussion prompts. There are ten chapters that we cover over two semesters. I conduct one (approximately) 40-minute long discussion (in German) per unit/chapter. The topic of the discussion is related to the theme of the chapter we have covered in *Kaleidoskop*. The discussion prompts are listed in the syllabus at the beginning of the semester, so students know well in advance what they will be discussing. The rubric is also available to students in Moodle. Students are expected to use relevant vocabulary found in the current chapter as well as particular grammatical forms we have discussed (verb tenses, adjective endings, subordinate clauses, etc.). See Table 4-22 for the prompts the learners use. Each of the ten discussion prompts are accompanied with these instructions: You may prepare for the discussion by writing your responses to the prompt. However, you may not read what you have written during our discussion. You may only discuss your response. Be prepared to ask questions of your classmates concerning their comments and opinions, and be prepared to answer their questions about your comments and opinions.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 4-22: *Prompts the learners use in German*

1.	*Sie sind „die junge Frau" in der Geschichte. Was sagen Sie Ihrer Freundin/Bekanntin, wenn Sie sie noch einmal einen Brief schreiben? oder Sie sind „die Freundin/die Bekanntin" in der Geschichte. Was sagen Sie der jungen Frau, wenn Sie sie noch einmal einen Brief schreiben?*
2.	*In "Eine Postkarte für Herrn Altenkirch" hat die Erzählerin ihn nie eine Postkarte geschrieben. Wäre die Situation anders gewesen, wenn sie die Möglichkeit hätte Facebook, Twitter oder Email zu benutzen? Was halten Sie von der Verhältnis der Erzählerin und Herrn Altenkirch?*
3.	*In Good-bye Lenin will Alex die DDR wieder lebenig machen. Warum? Hatte er recht so was zu versuchen? Was bedeutet es "Vater und Mutter zu ehren"?*
4.	*Das Mädchen in den „Sieben Raben" fühlt sich ihren Brüdern verantwortlich. Inwiefern ist jemand seinen Geschwistern verantwortlich? Müssten Sie schon im Leben für jemanden verantwortlich sein?*
5.	*Ist „Der Erlkönig" unheimlich? Warum oder warum nicht? Haben Sie schon eine unheimliche Situation erlebt? Wie haben Sie darauf reagiert?*
6.	*Beschreiben Sie eine ideale Firma oder einen idealen Arbeitgeber. Beschreiben Sie die ideale Karriere. Warum ist diese Karriere Ihnen ideal? Welche Kriterien oder Maßstäbe sind Ihnen wichtig?*
7.	*Wie kann ein Immigrantenkind erfolgreich sein? Wie können ihm seine Mitschüler hilfreich sein?*
8.	*Was für Probleme hatten Sie mit dem Einleben an der Universität? Inwiefern haben Sie jetzt mehr Freiheit? Haben Sie zu viel Freiheit? Vergleichen Sie Ihre Meinungen mit den anderen Kursteilnehmer/innen.*
9.	*Hatten Sie schon Heimweh? Wann, wo, warum? Wie kann man gegen Heimweh kämpfen?*
10.	*Was halten Sie von Recycling? Nehmen Sie teil? Wie, wie oft, wann, usw.?*

See Table 4-23 for English translations of the discussion prompts.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 4-23: ***English translation of the discussion prompts***

1.	You are the young woman in the story. What would you say to your friend/acquaintance when you write her another letter? Or You are the friend/acquaintance in the story. What would you say to the young woman when you write her another letter?
2.	In the story, the narrator never wrote a postcard to her former landlord. Would the situation be different if she had had the opportunity to use Facebook, Twitter or E-Mail? What does your response say about the relationship between the narrator and her landlord?
3.	In *Good-bye Lenin*, Alex attempts to bring back the DDR. Why does he do this? Was he right to do what he did? What does it mean to "honor your father and mother"?
4.	The girl in *The Seven Ravens* feels responsible for the fate of her brothers. To what extent is a person responsible for his or her siblings? Have you ever been in a situation where you were responsible for someone or something? Explain.
5.	Do you think *The Erlkönig* is eerie? Why or why not? Have you ever been in an eerie situation? Describe it. How did you react?
6.	Describe the ideal company you would like to work for. Describe your ideal career after graduation. Explain why this job would be ideal. What criteria are important to you?
7.	What would portend success for an immigrant child? What can his/her fellow classmates do to help him/her adjust to the new land and culture?
8.	What problems did you have in adjusting to life at the university? In what ways do you have more freedom? Can you have too much freedom? Compare your responses to those of your classmates.
9.	When have you experienced homesickness? What can one do to successfully combat homesickness?
10.	What is your opinion of recycling? Do you participate? Explain.

Test process. Students are aware of the schedule established from the beginning of the semester in the course syllabus. They are encouraged to work with their classmates in advance of the assigned discussion day to practice speaking about the topic. Group practice will not only help them understand the topic and related vocabulary, but will also help them understand one another's pronunciation. On the day assigned for discussion, we set our chairs in a circle. I pose the question in the prompt and discussion proceeds. The goal is to hold a true discussion, following any tangential ideas and eventually returning to the original topic, for 20-40 minutes. I

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

make notes based on the rubric and add leading comments and questions only as necessary.

Following the discussion, I review my notes and score each individual student according to the rubric, adding comments and suggestions for improvement as necessary. The rubrics are then posted on the Moodle course site for individual student review.

Contributor's Responses to the Questionnaire

Section One: Test Planning and Writing

Why did you write the test? What were the purposes of the test? Consider: To award a course grade, or part of a grade? Something else?
At the intermediate level of language proficiency (as measured by the ACTFL Guidelines for Language Proficiency, American Council on the Teaching of Foreign Languages, 2012a), students should be able to demonstrate proficiency in the four language skills (reading, writing, speaking, and listening). Testing for speaking proficiency meets a course objective. Intermediate language courses at our institution use the same objectives regarding language proficiency.

My rubric was designed as a means of evaluating the spoken discourse of intermediate and advanced level language students. Evaluating spoken language is difficult and subjective, in that pronunciation, grammar, and affective appearance influences evaluations. As part of awarding a grade for spoken performance in the intermediate level of German, it was necessary to develop a rubric so as to moderate these superficial influences in order to evaluate true linguistic proficiency, that is, the ability to speak to and to participate in conversation about a topic known to all concerned.

Did you write as many items or test tasks as you needed, or did you write more than you needed, then pare the number down? How did you decide which items or test tasks to keep? How did you decide which items or test tasks to discard?
The discussion topics are assigned from the beginning of the semester in the syllabus. Students know what the topic will be and can prepare. As we progress through the material in the textbook, I change or modify the speaking prompt so as to address related topics that arise as the students work through various texts and begin to make comparisons to other texts they have read. The final discussion topic is set at least two days before the discussion for evaluation is scheduled. This helps to establish content validity (Moskal & Leydens, 2000).

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you have one version of your test, or did you create a second equivalent version?
As discussion is fluid and cannot be repeated, there is no need to create other versions.

Were you concerned at how long the test would take to administer?
I allot one 50-minute class period to our group discussion. The question of time is a concern when the group is larger than ten. While I try to moderate each discussion, I try not to allow the discussion to be teacher-led. This is a problem at the beginning of fall semester as students have never participated in discussions which are evaluated. They expect the time to be a question and answer session. They readily respond to me, but are hesitant to pose their own questions or state their own opinions. In the situation, however, when certain students are verbose and "take over," it is necessary for me to redirect the discussion and to directly address students who are not participating.

The amount of participation is important as it addresses the maxim of quantity. Some students, however, are shy or inhibited when speaking German and need to be reminded that this is for an important course grade. Student hesitation to participate can affect construct validity (Moskal & Leydens, 2000).

Were you concerned at how long the test would take to score?
The rubric is extremely helpful with keeping track of the elements of a good conversation (the four maxims). I am able to make notes and use my own shorthand to record my impressions as they happen. I make an effort to score and annotate the final rubric for all students within 24 hours.

Were you concerned how you might use the test items themselves for learner feedback?
The rubric has space for my comments, which I try to make detailed. Setting appointments with individual students is sometimes necessary to point out ways to improve their participation in the discussion and to explain my expectations.

Did you consider making the test an open book test?
Students are allowed to have a note card with vocabulary words to which they may refer during the discussion. I also encourage students to get together to go over the material before the class discussion takes place, in order to be sure they understand the topic. I do not want learners to be completely blindsided should the discussion turn in an unexpected direction.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you consider allowing learners to use additional sources such as dictionaries, or their notes, while taking the test?
A notecard with vocabulary words is allowed.

Did you plan to allow learners to re-take a test for improvement? The same test, or a different test?
If a student must be absent and makes that known in advance, the discussion is scheduled for a different day. If a student misses the discussion, he/she has the option to discuss one-on-one with me.

What sources did you draw from for your test items?
Consider: Test item ideas or content from a textbook?
As noted previously, the discussion topics are based on the material covered in the course textbook and are modified according to how we addressed the material as we worked through it.

Did you consider learners' communicative competence when writing test items? What aspects of communicative competence?
I modify discussion prompts available in the course textbook so as to match students' communicative competence. For example, as the year progresses, I modify the prompts to include more critical thinking, which students are better able to handle as their affective filter lowers. They feel more comfortable with me and with their classmates.

I also grade more leniently during the first several discussions in the course than I do in later discussions. Learners are anxious early in the semester and this can keep students from demonstrating their full potential. However, by the middle of the first semester, students should feel comfortable with each other and with the material. I encourage them to push themselves to use more advanced vocabulary and grammatical forms.

What types of learner knowledge do you believe you are capturing in your test?
As we are a faith-based institution, students often include perspectives from their faith that may not be addressed in the material covered in class. The discussion allows them the opportunity to express opinions based in their own worldview. This requires them, as they prepare, to learn vocabulary that they might not have learned before.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

What learner skills do you believe you are capturing in your test? How does that change with test item types you used on the test?
The skill addressed in the assessment is speaking. While students speak in German in every class session, they may often be parroting what they read or heard, or they may be simply responding to comprehension-type questions. In the discussion assessment, they have to take their skill of speaking in German to the level of expressing an opinion.

If you wrote a performance test (where learners have to converse or present a topic, or write something above the sentence level), how did you get the ideas for what task learners had to do for the test?
Consider: From tasks learners do in class?
From tasks they have do to in real life?
We discuss topics and themes covered in class and in homework assignments. However, by extending the themes through prompts that require critical thinking, students have the opportunity to experience real-world language use which requires them to express opinions and consider the opinions of others.

How did you get ideas on how to score learners' performances (the scoring criteria)?
This is the most valuable part of the assessment. I teach a course in introductory linguistics. While reviewing Grices's Conversational Maxims (Grice, 1975), it occurred to me that these maxims could provide a means of evaluating student discussions in a manner considered to be related to real-world situations and not based on grammar, vocabulary, and pronunciation. The four maxims are quantity, quality (truth), relevance, and manner.

If you used points on a scale for scoring learners' performances, how did you decide on how many points on a scale to use?
Each maxim is graded on a five point scale. I prefer five points. Ten points would be too many and I tend to grade high, so I would not likely give grades in the middle.

Were you concerned about whether you could get a colleague to help you score learners' task performances?
Unfortunately, I am the only German professor at our institution, so I have no one who might help me score the discussions. Also, in order to have a second rater, that person would have to be available for the full hour of the discussion/test. I score the discussions alone.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you plan to give the scoring criteria to the learners for future self- or peer-assessment? Did you make another, perhaps simpler or shorter, version of the scoring criteria for learners to use?
I have not done this. However, it would be an excellent idea to divide the larger group into two groups and have each of the smaller groups assess the other. Another option would be to video record the discussion, make it available to all students and assign a self-evaluation or a peer evaluation as an assignment.

Did you compare your test to the textbook or other materials learners used?
The test is based on the material in their lessons. While the textbook has ancillary materials for the instructor, it does not include a test of spoken German.

Are you required to use specific tests in your program?
No. However, the Ohio Department of Education has available rubrics for spoken foreign language at the intermediate level. The criteria included are 1.) Being understood by a sympathetic listener (generally focusing on form and vocabulary); 2.) Convey and possibly extending the message; and 3.) Demonstrating intercultural competence. See:
http://education.ohio.gov/getattachment/Topics/Learning-in-Ohio/Foreign-Language/World-Languages-Model-Curriculum/World-Languages-Model-Curriculum-Framework/Instructional-Strategies/Scoring-Guidelines-for-World-Languages/Intermediate_Low_Holistic_Rubric_Presentational_feb2018.pdf.aspx?lang=en-US
While this rubric is useful, I see the assessment based on Grice's maxim as being more authentic.

Section Two: Test Administering and Scoring

Were you concerned about test security? What did you do to ensure test security?
For the purpose of this test, security was not a problem. Students had access to the speaking prompts from the beginning of the semester.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Were your tests photocopied, or did learners see your test items or test tasks another way, such as on the blackboard?
The speaking prompts were adapted from the course textbook. The prompts were either directly assigned from the textbook or were modified. Modified prompts were both written on the board and emailed to the students.

How did you deal with learners who missed the test, or who were late for the test?
If a student gave prior notification of an excused absence, I adjusted the schedule so as to hold the discussion on a day when all students were going to be present. Learners who missed the in-class discussion without prior notification, and whose absence was unexcused, received a zero for the assignment. In fall semester, six discussion grades make up 10% of the final course grade. In spring semester, four discussion grades make up 10% of the final course grade. If an emergency kept a student from participating in the group discussion, I would either not count the missed grade or set up a one-on-one discussion with me. This last option is not ideal as the purpose of the discussion is to get away from teacher-centered talk.

How did you prepare learners to take the test? Were there any test items or test procedures that were new to them?
Dealing with contingencies (van Lier, 1996) is always a concern in the fluent oral production of a new language. While all contingencies can never be controlled for, regular focus on vocabulary related to the topic of discussion and regular opportunities to speak German in the classroom help prepare students to construct comprehensible sentences and to overcome the inhibitions held by all language learners (Krashen, 1982).

The first discussion is always difficult as students expect a teacher-centered discussion. I have to balance the posing of questions in order to move the conversation along with waiting for someone else to pose a question or continue the conversation. If their inability or lack of desire to carry the conversation persists after the second discussion, I have to explain to them what I expect and remind them how important the discussion grade is to the final course grade.

Did you pilot your test? Do a trial run? Did the pilot result in any changes to the final version of your test?
I began using this assessment rubric for oral language production four years ago. Each semester is new and requires adjustments to the rubric. The current rubric is based on eight semesters of use in the intermediate German classroom.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you write any of the test in the learners' first language? Why?
The rubric itself is written in English (see Table 4-22). Students have access to the rubric before discussions take place. All speaking prompts are written in German in order to maintain the authenticity of the discussion (see Table 4-22).

For classroom tests: How did you accomplish scoring learners' tests?
I am the only German professor at our university; thus, I am the only person qualified to score the discussions. Ideally, this rubric would be scored by two people as the discussion takes place. Ideally, one scorer would be "a fly on the wall" and not participate in the discussion. I am considering the possibility of setting up a camera so that I can evaluate later.

Did you go back and change your marks on previously scored tests in response to problems you found while scoring tests later in the process?
Yes. I have sometimes reconsidered, after the discussion is complete, a mark on the rubric that I made early in the discussion. For example, a comment made by a student might not at first appear to be related to the topic, but as the discussion progresses, the relevance of the remark may become clear.

Did you later change your scoring because you recognized some learners couldn't answer certain test items due to some aspect of the items or the test?
Similar to the previous response, as I review my marks and comments, I sometimes see that I might have misinterpreted an early utterance which was clarified or corrected later in the discussion.

Did you put learners' responses to items into a spreadsheet for further analysis? Did that process help you catch scoring accuracy problems? Or problems with bias?
No. It might be a good idea to put learners' responses into a spreadsheet for further analysis as an idea for future evaluation.

Did you ask the students themselves to score their own test? Or a classmate's test?
Another option for scoring and for student learning would be to have students score one another. I might consider this for the spring semester. I could divide the class into groups of two or three and ask students not included in a group to evaluate the others. This could lead to high scores for students as students often give their classmates many benefits of the doubt.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Another option would be to record the discussion, as mentioned earlier, and have the students score their own performances.

For performance tests: How did you accomplish scoring learners' performances?
I score as the discussion takes place. The discussion takes place in a 50-minute period. I use pluses (+) and minuses (-) next to students initials to mark the four criteria of the rubric. I rescore before I prepare the final rubric and final score.

Do you think your test was reliable? What did you do to check?
There is a modicum of test-retest reliability, since I use the same rubric with the same students throughout the semester. Rubric scores for individual students have shown no great fluctuations in range, but for most students show steady progress/improvement. I would say that my test is reliable, because student scores have never varied significantly (Moskal & Leydens, 2000).

What did you do to ensure all learners had the same conditions under which they took the test?
All students are notified of the speaking prompt in the same manner and all students participate together in the discussion. Reminders of my expectations occur in the presence of all students.

Reporting scores

Did you report the scores to learners? If so, how did you report test scores to learners?
Scored rubrics, including comments in each of the four criteria on the rubric are posted to Moodle within two to three days of the discussion.

What was your goal in reporting the test scores to learners?
The goal, of course, is to help students see their problem areas and to direct them towards improvement. Nevertheless, calling attention to grammar or vocabulary deficiencies can often work to raise the affective filter (Krashen, 1982) rather than to lower it. And, despite notification of which areas they need improvement in, conversation/discussion proficiency can never really be controlled. Practice is what makes perfect.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you spend time explaining scores, or answering learners' questions about scores in or out of class?
I am always available to consult with students about their scores. Students rarely ask about them.

Did you report peer-assessment scores or self-assessment scores on the test?
In the future, if I decide to conduct peer assessment, I will most likely share the scores, but not calculate them as official scores.

How quickly did you report scores to learners? Was speed a priority?
As with all other scores, I seek to report them within two to three days, a week at the most.

Section Three: Using Test Scores

Cut scores

Did a colleague or supervisor suggest cut scores?
No. This test is a test of oral performance. Students are assigned a speaking prompt and are evaluated with the rubric. During fall semester, students are evaluated six times; during spring semester, this evaluation occurs four times. The six or four scores combined equal 10% of the final grade for the course. The course grades/letter grades are based on ten-point spreads with 90-100 being an A, etc. The rubric outcomes are each evaluated from one to five points with five being the highest possible. The highest possible score for the evaluation is 20 (four criteria).

Did learners' test scores have any positive or negative consequences for you, in terms of your institution?

No. I am allowed to determine my own evaluations and how scores from those evaluations will be used to determine the final course grade.

What was the role of the test score in determining learners' grades? Consider: How much weight did you give your test? How did you decide? Were other measures used to decide learners' grades, besides your test?
The scores from this evaluation are for my use only. They are used as part of the full course evaluation for Intermediate German I and Intermediate German II (college level). This oral evaluation makes up 10% of the final course grade:

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Attendance and Participation	10%
Aufgaben:	20%
(Textbook Exercises, Workbook Exercises, Lab Manual Exercises—written and spoken)	
Diskussionen	***10%***
Aufsätze	15%
Portfolio	10%
Tests	25%
Final Exam	10%

Did your test capture some knowledge, skill, or ability the other measures did not capture?
Testing oral proficiency is difficult and particularly difficult to evaluate. I use the discussion prompts provided in the course textbook, but needed a valid measure for evaluating students' performance (Moskal & Leydens, 2000).

Instead of testing oral proficiency in a one-on-one situation as many professors do, I decided a group discussion would be more authentic, requiring students to negotiate meaning, to make themselves comprehensible and to negotiate listening (van Lier, 1996). The use of Grice's four conversational maxims seemed to be a reasonable way of designing the criteria for the rubric (Jonsson & Svingby, 2007).

Reporting scores

How did you report scores to learners? Was timeliness of concern to you?
I am always concerned with timeliness. I try to have scores on every evaluation to students within one week's time, sooner if possible. For the discussion evaluation in question here, not only is it important for students to have their scores in a timely manner, it is important for me to complete the rubrics within 24 hours, as there is a subjective component to scoring that does not appear on the rubric itself. I have to remember the discussion.

Did you hand the test back to learners? Did the learners get to keep the tests? Or did you take the tests back?
The rubrics with scores and comments are posted to Moodle. Students have access to them there.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you offer feedback to individual learners in addition to their test scores? Written? Orally? In or out of class?
The rubrics available to the students provide written feedback. In addition, I am always available for feedback outside of class. However, no student has ever asked for assistance in bettering their scores.

Did you teach learners to interpret their scores?
Scoring on the rubric is straightforward and the weight of scores is explained in the syllabus.

For performance tests, did you use the test criteria to help learners interpret their scores?
The test criteria are explained on the rubric. A blank rubric is available to students either as part of the syllabus or on Moodle.

Using test scores

Do you feel you learned what you needed to learn from the test results about what learners could and could not do, or what they knew or did not know?
Yes. I feel that the rubric adequately provides a picture of the oral language the students are capable of producing (Jonsson & Svingby, 2007). After the first or second discussion (test) of the fall semester, I have to explain to learners that I want to be "a fly on the wall," while they as a group discuss the assigned topic. At this point, they are more familiar with each other and are able to lower the affective filter during discussion sessions (Krashen, 1982).

Did your test change how learners studied?
Consider: Did you use particular item types or a performance test to change learners' practices or support their learning?
Did their scores indicate they had changed their learning practices?
The discussion is assigned through the course syllabus; therefore, students know well in advance when the discussion will take place and what the topic of discussion will be. Nonetheless, they often wait until the night before the discussion to start prepping. I tell them in advance that they might want to practice with other classmates and compare notes on the topic. I also suggest they look up and learn vocabulary and grammatical forms that are pertinent to the topic. When I observe students use new vocabulary and push themselves to use more sophisticated grammar during the discussion, I know that this evaluation has changed their study habits.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you spend time going over the test in class?
I make sure that students understand the discussion prompt and explain how we might take the discussion further.

Did learners ask you about the test itself (not the test scores) outside of class?
I have been using this evaluation for three years and have not yet been questioned about the scoring.

Did learners' test scores change your teaching?
I try to make sure that classroom discussion of the topic to be evaluated is clearly understood by the students and that they begin to think about different interpretations of the topic under consideration.

If you could turn back time, what would you change about your test? What would you change about your test administration?
In section one or two, the question was raised as to whether I had recorded these discussions. I have not. I would like to make recording a part of each of the four discussions in spring semester. And as much as students may not want to hear or see themselves speaking German, I would like to work into the schedule opportunities for them to review their own performance, or to act as a peer reviewer for a classmate.

Section Four: Evaluating and Reviewing your Answers

Please read through your answers to the items and answer the following:
To what extent is your test here an innovation, or something new, for you?
My previous efforts at assessing oral production at the intermediate level of German did not yield satisfying results. Assessing for only comprehensible pronunciation and grammar did not provide a holistic review of how the student speaker might be received and understood by native speakers. The ability to pronounce perfectly and produce grammatically correct sentences does not portend comprehensibility of content. Similarly, comprehensibility of content, particularly when colored by intercultural understanding or misunderstanding, is indeed incomprehensible of the speaker's grammar and pronunciations are not clear. The balance, provided by an analytically holistic assessment, can provide the best assessment of the speaker's ability to produce language.

My rubric, based on Grice's four conversational maxims (Grice, 1975) is novel and assesses in an analytically holistic manner that is expected by listeners in day-to-day conversations. Because topics assigned for assessment in

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

the classroom require the students to generally stay on topic, modifying the maxims to include grammar and pronunciation was reasonable and easily done. I have used the rubric now for six semesters and find that it meets my needs and is relatively easy to use. Students are able to see where their contributions to the discussion are missing and generally work to improve the manner, quantity, quality and relation of their contributions/utterances. They work to include new vocabulary and concepts; they work cooperatively to take the discussion in directions not necessarily part of the original discussion topic. These discussions allow them to demonstrate their intercultural competence and to address how their faith influences their intercultural competence.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

A Multiliteracies-oriented Project-based Assessment for Intermediate Foreign Language Italian Classes

Borbala Gaspar
University of Arizona
& Margherita Berti
University of Arizona

Introduction

Borbala Gaspar is a Ph.D. candidate in Second Language Acquisition and Teaching (SLAT) at the University of Arizona and holds an M.A. in English as a Second Language. She has been teaching Italian for twelve years at all four undergraduate levels at the university level. Her main research interest includes learner agency development and imagination. Margherita Berti is also a doctoral student in SLAT at the University of Arizona and holds an M.A. in Linguistics/Teaching English as a Second Language. She teaches undergraduate Italian courses and has four years of experience teaching Italian, Spanish and ESL. Her research focuses on materials development for assessment and pedagogical purposes with the integration of technology tools for foreign language courses.

This multiliteracies-oriented test is currently being used in intermediate Italian courses at a large public university in the southwestern United States. It was designed to replace the traditional oral assessments that are question-answer based, during which learners are providing predictable short answers to predictable, predetermined questions. In the multiliteracies-oriented test presented here, students perform tasks designed using the pedagogy of multiliteracies (Cope & Kalantzis, 2009; Kern, 2000; New London Group, 1996). A multiliteracies framework perceives learning as a process of discovery (Paesani, Allen, & Dupuy, 2015, p. 23), and learners are not only preparing for an assessment, but they are also engaging in forethought, design, and reflection while reading, writing, listening, and speaking. A multiliteracies pedagogy integrates, rather than separates, the study of

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

language (the practice of all four skills) with the study of literary-cultural content (Paesani et al., 2015, p. 22). In this test learners are given autonomy and take an active part in their learning, while contributing to the content of the test since they select their own topic.

We consider our approach to be ecological. By ecological, we mean the ways in which we perceive the learner and the teacher in the educational context, as outlined by Van Lier (2004). The learner is seen as a whole person, not a grammar-production unit, someone who has meaningful things to do and say, and who is being given responsibility and taken seriously. The teacher, on the other hand, aids in a scaffolding process that focuses on the learners' interests and skills development (Van Lier, 2004, pp. 223-224). We also consider this test a holistic assessment which integrates all four language skills where students make sense of multimodal texts and resources. By holistic, we mean the complex connections that are made as an end result of learners engaging in projects and tests such as those outlined here. These connections are between the target language community, the classroom community, and the local community. During this process learners engage with multimodal texts and they reflect and compare different linguistic styles and modalities with their own linguistic choices (Warner, 2011, p. 7). Learners also become experts in their area, develop research skills, and teach their peers about a specific topic in the target language. In doing so, they make connections with language-use communities.

The test is a semester-long project consisting of five steps submitted through a university-sponsored online platform, D2L (2019, https://d2l.arizona.edu) and/or shared on Google Drive (2019, https://drive.google.com). In addition to the five steps, learners participate in four in-class workshops in which they scaffold, construct, shape and finalize their individual research projects (see Table 4-24). This is a mostly formative assessment designed to monitor students' learning through teacher and peer dialogue using both oral and written communication. The final step of the test, a final oral examination, is a summative assessment, and done for accountability purposes. Learners are graded in all the five steps of this assessment.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 4-24: ***Description of the four in-class workshops***

Workshop	Description
Workshop 1: Pre-speaking and pre-learning activity	Prior to the beginning of the research project, learners are asked to brainstorm keywords related to the topic they want to explore. In a "keyword cloud" students write all the Italian words that come to their mind in relation to the chosen topic. Students are also asked to take note of what they already know about their topic. Finally, students work in pairs to discuss and exchange information about their research project and provide at least two suggestions/comments to their partner.
Workshop 2: Interpretation of multimodal resources (multimodal refers to the combination of a digital text with one or more components such as sounds, images, words, etc.)	Students learn how to search for Italian resources on the Internet using Italian-language search engines. They learn how to search for additional keywords in a text and how to make sense of information found online. This workshop also addresses the use of social networking sites such as Instagram, Facebook, Twitter, etc. Learners collect keywords and resources that they will use for their research project and discuss their goals and primary findings.
Workshop 3: Making meaning of texts	As an at-home assignment, students select, read and write a 3-2-1 summary (students describe three takeaways, two questions, and one thing they enjoyed about a specific resource) of an article related to their research project. In class they present their article in Italian with the use of brief notes and pictures. This workshop aims at preparing students for the final oral presentation.
Workshop 4: Knowledge application and reflection	Learners present their findings in Italian to one peer in the classroom. They are also asked to share how they plan to engage the audience during the final oral presentation as well as to give instant feedback to their partner.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

The Test

This is what students receive at the beginning of the semester. The test is a semester-long research project consisting of five phases:

1. *Research proposal draft*
   The research proposal draft must be 2 to 3 pages and it is your plan of research on a pre-approved topic of your choice. The following sections must be included:

   a. Introduction: Explanation of the reasons why you selected your topic and how you connect to it (personal and professional connections).
   b. Research questions: Specific questions that address particular details of your chosen topic. At least five research questions must be included. You should not be already familiar with the selected topic, rather you should be exploring something that is new and interesting to you.
   c. Resources and evaluation: List of resources that you plan to use to answer your research questions. At least ten resources must be used. Resources can include videos, websites, social networks, books, magazines, etc. For each resource explain how it will help you answer your research questions.
   d. Findings: A brief description of what you have found based on your initial research on the chosen topic.
   e. Goal: What you hope to achieve with this research project. Reflect on the implications for yourself and your classmates.
   f. Class-engaging activities: Explanation of how you are planning to engage your peers throughout your presentation. Think about tools that you can bring in the classroom such as handouts, visuals, brief activities, group quizzes, etc.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

g. Community outreach: Connection between your research topic and the wider community, whether local or global.

For the scoring rubric for this step, see Table 4-24.

2. *Revised research proposal*
After your instructor has provided you with feedback on your research proposal draft, please make the appropriate changes and create a finalized copy of your research proposal. This assignment is scored using the same rubric of the research proposal draft.

3. *Draft of presentation slides*
In the draft version your presentation slides you must include, at least, the following sections:

a. Introduction: You illustrate the connection between the chosen topic and yourself, the presenter, and draw the audience into the presentation in creative ways, for example with well-formulated questions, graphics and sounds, brief quizzes, etc.
b. Content: You describe synthesized research and show critical thinking skills. Your resources must be meaningful and linked to your research topic.
c. Class-engaging activities: You must include activities that allow for interactions with your peers. Such activities are not simply in the question-answer format, rather they encourage reflection and collaboration.
d. Findings: You synthesize the outcomes and findings of your research project. This section must be clear and should enrich the audience with new knowledge.
e. References: Include links to all the resources that you are using in your project in the final slide(s).

For the scoring rubric for this step, see Table 4-26.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

4. *Revised presentation slides*
   After your instructor has provided you with feedback on your research proposal draft, please make the appropriate changes and create a finalized copy of your research proposal. This assignment is scored using the same rubric of the presentation slides draft.

5. *Ten-Minute oral presentation*
   At the end of the semester you will present your semester-long research project to your classmates during our regular class time. Please ensure that your slides are completed and ready for you to present. During the presentation make sure to engage your peers, keep eye contact, refer to your screen sporadically and ensure that your voice is being heard. This is the time to share with everyone what you have been working on this semester.

For the scoring rubric for this step, see Table 4-26.

Rubrics for Scoring

Table 4-25: ***Research proposal draft (Step 1)***

_____/35 points	5-4 points	3-2 points	1-0 point
Introduction	Well-formulated introduction that includes a clear explanation of the importance of the research topic and provides connection to personal interest.	General introduction includes explanation of the importance of the research topic. There is a general connection to personal interest.	Introduction lacks focus, there is no clear explanation why the student selected the topic. The connection between the project and the student is missing.
Research questions and evaluation	Very clearly stated questions that are specific and to the point. Questions are promising in resulting of original and interesting research findings.	Clearly stated questions, however they are rather general and not specifically focused.	Less than five questions, which are general and not focused. Lack of details in the questions.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Resources	Resources are versatile (videos, articles, blogs, social media, etc.). Each resource is for a different and/or similar aspect and it is linked to the research questions. The student started to critically evaluate resources and paired them to the research questions.	Resources are rather limited and partially related to the research questions. It is somehow understandable how resources will be used to answer the research questions.	Less than ten resources. Research questions do not seem related to the selected resources. Resources are confusing and not critically selected for the project.
Findings	Initial, specific and new findings are clearly outlined.	Initial findings are outlined but rather general.	Initial findings are not clearly explained.
Goals	Goals for this research project are clearly stipulated.	Goals for this research project are described, but are rather general.	Goals for this research project are vague and not clear.
Class-engaging activities	It is clear how the student plans to engage others throughout the presentation and peers will acquire new knowledge. It is clearly explained what tools will be used during the presentation. It is specified if the student will bring items and/or tools the day of the presentation. It is clear how such items and/or tools will be utilized.	It is somehow understandable how the student will engage others. There is no explanation of what tools will be used during the presentation.	No engaging activities are included in the research proposal.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Community outreach	In the proposal the student strongly considers how the project connects to communities, whether local or global.	The project includes a community outreach section; however, the connection is weak or not elaborated.	The project includes a community outreach section; however, it is not well-planned, or parts are missing.

Table 4-26: ***Presentation slides draft (Step 3)***

/50 points	5-4 points	3-2 points	1-0 points
Introduction	The introduction connects the presenter with the chosen topic and sparks interest. It draws the audience into the presentation with well-formulated questions or other creative ways.	The introduction somehow connects to the presenter and partially engages the audience with some superficial questions.	The introduction is basic and it is not clear how it connects with the presenter. The audience is not involved in the introduction.
Use of primary resources	In the slides it is clear that the student synthesized research. The slides also show critical thinking skills and all resources are logically connected to each other (not a list of information).	In the slides it appears that the student partially synthesized research. Although critical thinking is not evident, the student somehow connected resources together. The use of quotes is minimal and all resources are listed on the references slide.	In the slides it appears that the student copied and pasted information from the resources or just added information without synthesizing it. Student did not show critical thinking skills and the content is disconnected with

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

	The use of quotes is minimal and all resources are listed on the final references slide.		manyunnecessary quotes. Some resources are missing on the references slide.
Content depth and transitions	The topic is covered extensively and includes many details. It uses advanced vocabulary from the course and from external resources. Tools are used creatively and extensively. The information provided flows well from a slide to the next one.	The topic is sometimes covered throughout, and other times covered superficially. Some of the vocabulary from the course is used. It seems that some information is lacking.	Only a few details are provided with basic and not varied vocabulary. The content presented looks like a "grocery list" with no depth or connections.
Content accuracy and comprehensibility	There is plenty of supporting information, evidence, images, etc., to make the presenter's point. Content is clear and easy to understand.	There is a fair amount of supporting information, but it is somewhat sparse. The presenter sufficiently provided content in support of the topic. There are some parts that are not easy to understand.	The content is presented superficially and it does not seem accurate. The presenter did not provide content in support of the topic. There are many parts that are difficult to understand.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Originality and creativity	Originality and creativity are clearly visible. The student used attractive and meaningful pictures or other creative tools to present the topic. Fonts and colors are well chosen to reflect the presenter's purpose and support the presentation.	Originality and creativity of the student are somewhat visible. The student used fonts, colors and other creative tools to present the topic. However, they are sometimes inconsistent and at times they do not support the presentation.	Originality and creativity are lacking. The student does not show any originality or creativity. Slides are basic and superficial.
Pictures, videos and background choices	Pictures and videos are labeled and are well integrated. They have a specific meaning and/or function in the presentation. The background enhances the presentation.	Pictures, videos and background may not be distracting from the content; however, they do not enhance the presentation and comprehensibility. Some pictures are just decorative.	The layout and color choices distract from the content of the presentation. Some of the images are purely decorative and seem out of place.
Clarity	The amount of content is not overwhelming. The presentation flows logically, it is easy to see and read.	Some parts of the content are not well balanced: there is too much or not enough information. There are either parts that are unclear or hard to read.	The content is unclear. The information presented is very hard to read and incomprehensible.
Engaging resources and activities for peers	The presenter included engaging resources and tools to interact with classmates. Engaging activities include	The presenter included only a few engaging activities, although they are mostly in a question-answer format without collaboration and reflection	No engaging activities and/or very minimal resources are included in the slides.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

	reflection and collaboration opportunities for all peers throughout the presentation.	opportunities. The presenter did not engage creatively all peers throughout the presentation.	
References slide (primary resources)	Contains at least twelve resources from Italian-language websites. These resources are used throughout the presentation and it is clear that they were selected for a purpose. Selected resources show an attempt to provide well-developed research.	Contains at least ten resources from Italian-language websites. These resources are used throughout the presentation. It is clear that most of these resources were selected for a purpose. Selected resources do not always show an attempt to provide well-developed research.	Contains less than ten resources. Not all the resources are used throughout the presentation. Selected resources do not show an attempt to provide well-developed research.
Knowledge gained and findings	The findings have a clear focus and there is a take-home message. It is a great synopsis of personal research; it is clear that the presenter spent a lot of time researching the topic. The slides are organized and presented in a manner that facilitates learning for all members of the classroom. There is a	Presentation loses focus at times, providing a lot of information on different areas. The take-home message is not clear. There is a good synopsis that shows that the presenter did research the topic. The slides are organized and presented in a manner that in most cases facilitates learning for all members of the classroom. There is a weak connection to the	It is not clear how the findings relate to the presentation. There is no take-home message and the presenter did not research the topic. The slides are not organized and do not facilitate learning. There is no connection to any community.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

	strong connection to the local or global community.	local or global community.	

Table 4-27: ***Oral presentation rubric (Step 5)***

/100 points	10-8 points	7-5 points	4-0 points
Eye contact	Maintains a continuous connection with the audience through eye contact and gestures.	Maintains a connection with the audience most of the time through eye contact and gestures.	Maintains a minimal connection with the audience through eye contact and gestures.
Position	Does not hide behind the computer during the presentation; moves and turns toward the audience during the presentation.	Most of the time does not hide behind the computer during the presentation and moves and turns toward the audience during the presentation.	Hides behind the computer and/or does not move toward the audience during the presentation.
Presentation mode	Talks clearly, with good pronunciation, loud and not very fast or slow.	Most of the time talks clearly, with good pronunciation. Some parts are either fast or slow.	Most of the time the speech is unclear and hard to follow.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Engaging activities	Engages classmates with meaningful activities and focused questions.	Interacts with classmates through superficial activities and/or questions.	Minimal interactions with the audience.
Verbal communication with audience	Able to demonstrate spontaneous communicative competence in Italian.	Able to demonstrate spontaneous communicative competence in Italian most of the time.	Unable to demonstrate spontaneous communicative competence in Italian.
Preparation	Presenter is in charge of the presentation. Controls and manages well the time, it is clear that this presentation was practiced several times. Glances occasionally at notes but does not read them.	Most of the time the presenter is in charge of the presentation. There are some difficulties with time management and/or often reads the notes.	The presenter seems to lose control of the presentation. There are difficulties with time management and/or reads most of the content from the notes.
Visual aid	Continuously uses images and content to point out, show, and clarify the presentation.	Most of the time uses images and content to point out, show, and clarify the presentation.	Rarely uses images and content to point out, show, and clarify the presentation.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Critical thinking	Presentation is not a report on facts, instead it has a clear take-home message that is original, based on the findings of the presenter. Presentation is thought-provoking and inspiring. Presenter sparks interest from the beginning and clearly explains the findings.	Presentation provides a lot of information on different areas, but the take-home message is not clear. There is a good synopsis which shows that the presenter did research the topic but lacks critical engagement and/or lacks in sparking interest.	Presentation is not logically organized. It is rather a summary than a synopsis of personal research. It includes very general information and/or unsuccessful in sparking interest.
Comprehensibility	Presentation is easily understood. Verbal tenses are correct and complex sentence structures are used.	Presentation is easily understood. There are some obvious errors in sentence formation and verbal tenses which impede comprehension. Somewhat complex sentence structures.	Hard to understand, very basic and short sentence structures. Multiple errors in sentence formation.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Organization	There is a clear logic and sequence of how information is organized, presented, and how the presenter structured the findings and conclusion.	Sometimes there is no clear connection on how the presenter structured the findings and conclusion.	There is no clear connection on how the presenter structured the findings and conclusion.

Contributors' Questionnaire Responses

Section One: Test Planning and Writing

Why did you write the test? What were the purposes of the test?
An educational system that prepares learners to be competent global citizens, who can communicate effectively in more than one language, is essential in today's society. The 2007 Modern Language Association (MLA) report calls for an approach that encourages transcultural competence and provides learners with tools that foster reflections about the world and helps them see through a critical global lens and not just their own (MLA Ad Hoc Committee on Foreign Languages, 2007). Keeping in mind the 2007 MLA report, we wrote this test with the aim of exploring new ways to assess and engage students in our intermediate Italian classrooms.

This test differs from a traditional question-answer based oral test (e.g., the instructor asks: *What did you do yesterday?* student responds: *I went to the movies*). We wanted to assess students more holistically (assess reading, writing, listening, and speaking), promote critical-thinking skills, encourage agency, autonomy and motivation, by giving students the opportunity to explore the content they prefer and by providing them continuous guiding and support.

By letting learners choose the content we increased their interest and we fostered opportunities to connect their personal and/or career interests with the Italian language. Since the test had several components spread out during the semester, students were given continuous feedback (e.g., instructor feedback, peer feedback) and guidelines as they constructed their project. This scaffolding process incremented the depth of learners' engagement and learning. Boblett (2012) describes scaffolding as "a system

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

of temporary guidance offered to the learner by the teacher, jointly co-constructed, and then removed when the learner no longer needs it" (p. 1), and this is what we aimed to do with this test. This project-based oral assessment fosters critical reflections, increases focus on the construction of meaning in different genres (whether it is spoken or written discourse), and promotes active and personalized learning.

At the core of the design of this project-based test is the multiliteracies framework that views learning "as a dynamic process of discovering form-meaning connections through the acts of interpreting and creating written, oral, visual, audiovisual, and digital texts" (Paesani et al., 2016, p. 23). The multiliteracies model is divided into four pedagogical stages: situated practice, overt instruction, critical framing, and transformed practice. Through these stages we engaged students with literacy and helped them make sense of communication, which occurs through graphics, music, sound, print, images. The design of our test also aligns with Van Lier's (2004) ecological approach to language learning that sees the learner as a "whole" person, not as a grammar production unit. Van Lier's ecological approach also considers the teacher an entity that gives assistance to the learners so that they can develop their skills.

While we believe that proficiency is an important part of language courses, we also wanted to balance the development of skills and give students the opportunity to think critically, express themselves meaningfully and give them responsibility and control of their learning.

How did you decide how many subtests to write?

Subtests were created on the necessity to develop a well-planned and organized test that follows the multiliteracies framework. We used an assessment model described by Paesani et al. (2015, p. 129) which recommends a four-stage lesson plan that includes pre-speaking activities, textual interpretation, knowledge transformation and conclusion, and overall reflection. We divided our test into five phases with the following components:

Subtest 1 is the research proposal draft (see Table 4-25). Students develop their research proposal draft using the rubric as a guideline and the instructor uses the same rubric to give a grade and suggestions on how to improve the proposal. Learners are given scores and feedback through the instructor's comments on the proposal draft, suggesting additional resources and fostering critical reflections.

Subtest 2 is the revised research proposal (see Table 4-25). Students use the feedback received on the research proposal draft to improve and further develop their research. Learners hand in the revised version of the proposal

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

based on the feedback, comments, and guidelines received from the instructor. The revised research proposal is evaluated using the same rubric of the draft submission.

Subtest 3 is the presentation slides draft (Table 4-26). Here learners participate in a good/bad slide activity, where students work in pairs, and evaluate and grade two slide presentations. Students then develop their own slides at home and submit their slides online. The instructor's feedback is given directly on the slides and a score is assigned using the rubric.

Subtest 4 is the revised presentation slides (Table 4-26). Students use the suggestions received on the presentation slides draft submission to improve and finalize their slides. The revised version is evaluated using the same rubric of the draft submission.

Subtest 5 is the final oral exam presentation (Table 4-27). Before the final oral presentation, students work in pairs, rehearse their presentation, and give feedback and a grade to each other based on the rubric. This activity is for practice purposes and helps students prepare for their final presentation. On the day of the presentation, learners are graded using the rubric and they are given a score and feedback on D2L.

Did you have one version of your test, or did you create a second equivalent version?

There was one version of this test. We did however accommodate students' needs. For example, those who had anxiety issues were able to complete the last subtest (the classroom presentation) by recording themselves presenting their project outside of the classroom setting, without a live audience.

Were you concerned at how long the test would take to administer?

Overall, we were concerned about the time this test would require in class as well as out of class. We already had a preplanned agenda with specific topics and grammar structures to be covered over the semester. While we anticipated that a semester-long research project divided into five phases would be demanding, we were able to successfully incorporate the project into the curriculum.

We incorporated certain presentations in the curriculum, such as a project on Italian coffee. The instructor collaborated with the presenter and created an activity where students put in order the steps for preparing coffee using the passive voice, a topic in the curriculum.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Were you concerned at how long the test would take to score?
We were mostly concerned about how to design the test in a way that was time efficient since the test was spread out throughout the semester and consisted of different parts. The scoring itself did not take a lot of time because the rubrics we created were very detailed and provided feedback to the learners.

Were you concerned how you might use the test items themselves for learner feedback?
This test scoring encouraged two-way feedback and collaboration between the students and the instructor. Learners were not passive but rather involved in their projects by asking questions and clarifications in line with the feedback and scores they received. After the research proposal draft (Table 4-25) and the presentation slides draft (Table 4-26) were submitted, the instructor provided detailed feedback and then students submitted a revised version of each subtest.

Did you consider having your students take your test on a computer?
This test was not meant to be taken at one single time on a computer since it was divided into steps completed at home. Nonetheless, this test required students to use technology because learning resources needed to be found online.

Did you consider making the test an open book test?
This test was an open book test in the sense that students needed to look for and make sense of resources around them, whether online or not, in order to complete the various subtests. The final subtest, the oral presentation, required students to use slides and notes, if needed.

Did you consider allowing learners to use additional sources such as dictionaries, or their notes, while taking the test?
While developing their projects learners were allowed to use additional sources. During the presentation of their findings (the fifth subtest) learners were asked not to read their slides word-by-word. They were, however, allowed to have notes.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you plan to allow learners to re-take a test for improvement?
Each subtest was designed to provide feedback and a grade to the learners. After receiving feedback, learners submitted a revised version, considering the instructor's comments. While students could improve the research proposal draft and the presentation slides draft with their revised submissions, the final oral presentation could not be re-taken.

What sources did you draw from for your test?
The test was prepared by taking into consideration the World-readiness Standards outlined by ACTFL (National Standards Collaborative Board, 2015), specifically, the five "C" goal areas. The ACTFL Standards stipulate that language learning can go beyond the classroom setting by utilizing resources that are around students, whether online or not, and by connecting with communities within and beyond the university.

Did you consider learners' communicative competence when writing the test? What aspects of communicative competence?
Yes, we considered learners' communicative competence when writing the test, especially since the culminating part of this project was an oral presentation in the target language. This test helped learners make sense of relevant texts and resources in the target language of a chosen topic and prepared them to report their project and talk about the findings to classmates. Learners had to engage in strategic competence by tailoring and modifying the resources (e.g., online articles, YouTube videos, social networking sites, webpages, etc.) to the linguistic level of the classmates. They engaged in discourse competence by connecting ideas in a coherent manner while preparing for the presentation. They engaged in linguistic competence by using appropriate language conventions during the presentation. Finally, students were able to answer questions about their topic and engaged classmates throughout the presentation with interactive activities.

Which test item formats do you prefer to use?
This test is a performance test since students were required to perform tasks, rather than just producing or selecting the correct answer. We consider this test a holistic assessment where students are not only making sense of multimodal texts and resources; they also become experts in their area, develop research skills, and teach their peers about a specific topic in the target language.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

What types of learner knowledge do you believe you are capturing in your test?
In this test learners are considered individuals with personal interests, with general and specific knowledge, and previous life experiences. Through their research, students put their knowledge into question and gain expertise in one particular area. Learners begin their project with general knowledge which becomes more specific and in-depth by the end of the semester.

What learner skills do you believe you are capturing in your test?
In this test we are capturing a holistic language competence in all areas: listening, reading, writing, and speaking. The test items required students to listen to videos, write notes, read online articles, and present in front of their classmates in the target language.

If you wrote a performance test (where learners have to converse or present a topic, or write something above the sentence level), how did you get the ideas for what task learners had to do for the test?
The tasks were chosen to reflect what learners do in real life. The goal of foreign language learning is to prepare learners to be able to use the language, whether it is speaking, listening, writing, or reading, in real-life situations. This test prepared learners to make sense of relevant multimodal texts, synthesize findings, write, speak, and plan creative activities in the target language. Additionally, students connected their topics with communities beyond the classroom walls. For instance, one learner investigated available food resources for university students in Italy. She then collaborated with five other students in the class to design and prepare recipe cards. Six different recipe cards were created specifically for university students and distributed at the Campus Pantry of our university.

How did you get ideas on how to score learners' performances (the scoring criteria)?
We created rubrics keeping in mind what we wanted learners to include in their assignment. An important goal was to make sure that learners researched their topic in depth. We developed three rubrics: One for the research proposal (Table 4-25), one for the presentation slides (Table 4-26), and the third one for the in-class oral presentation (Table 4-27). We shared the rubrics with students at the beginning of the semester so that they knew what we expected from them. The rubrics reflected elements of the test tasks. For example, in the presentation slides students had to do the following, among other things, "*Findings: you synthesize the outcomes and*

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

findings of your research project. This section must be clear and should enrich the audience with new knowledge."

If you used points on a scale for scoring learners' performances, how did you decide on how many points on a scale to use?
We wanted to emphasize that the work learners put into researching and creating their research project had overall more weight than the final oral presentation. The research proposal draft was worth up to 35 points, with each category within it worth up to 5 points. The presentation slides draft was worth up to 50 points, with each category worth up to 5 points. The oral presentation was worth up to 100 points, with each category worth up to 10 points. The choice to use a 1-5 scale and a 1-10 scale was dictated by the subsections (scoring criteria) in each rubric.

Were you concerned about whether you could get a colleague to help you score learners' task performances?
We were not generally concerned about scoring inconsistencies since the rubrics were very detailed and allowed us, as well as students, to know what was expected in this project.

Did you plan to give the scoring criteria to the learners for future self- or peer-assessment?
We gave the same scoring criteria to students. During an in-class activity, we asked students to work in pairs and grade two pre-made presentation slides (a good one and one that needed significant improvement) with the rubrics. The goal of the activity was to help students understand what we were looking for and how to develop presentation slides for a foreign language course.

Did you compare your test to the lessons that learners had?
Since this is a holistic test that includes all language skills, we think there were similarities to the lessons that learners had during the semester.

Are you required to use specific tests in your program?
Yes, we are required to administer specific tests at the end of each chapter.

Was your test different from required tests? How?
Yes. It was a multiliteracies-oriented project-based test which sees foreign language learning and assessment holistically and aims at the development of critical thinking skills, reflections, and connections to communities. It

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

differs from other tests since it does not focus on specific grammar structures.

Did you use items or ideas or content from previous tests?
Ideas were developed from a project-based test previously used in intermediate Italian courses by one of the authors.

How did you change the parts you decided to keep? Why?
We kept the test as project-based in the intermediate level since allowing students to choose a topic would increase their motivation and help understand the usefulness of the target language beyond the classroom setting.

What did you add that was new, or different? Why?
To this project we added the multiliteracies framework, workshops, and the proposal writing stage.

Section Two: Test Administering and Scoring

Were you concerned about test security?
One concern was that, being an open-book test, learners might use online translators such as Google Translate when writing or making sense of multimodal texts, instead of trying to figure out the meaning through learning strategies. A secondary concern was that learners might copy and paste sentences found on Italian websites in their presentation slides, instead of understanding the language.

However, based on our experience the abovementioned concerns were not noteworthy because learners read the same topic on different webpages and they became more and more familiar with the vocabulary which they used in their own presentations. Learners were aware of the syllabus rule that "writing three words next to each other as in a resource is considered plagiarism" therefore they made sure to work with the texts and resources they had in front of them. Additionally, the submission of the proposal draft and the slides draft allowed us to monitor for plagiarism.

How did you deal with learners who missed the test, or who were late for the test?
Learners were given specific deadlines throughout the semester to complete each subtest. Learners received only partial credit for test-related assignments turned in late.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

How did you prepare learners to take the test?
Learners participated in in-class workshops (see Table 4-24). The workshops really helped students understand how to respond to the various parts of the test.

Did you pilot your test?
Yes, this project had a pilot version in a previous semester.

Did you write any of the test in the learners' first language? Why?
The test task directions and the scoring rubrics were written in the learners' first language to maximize clarity. However, the in-class workshops (Table 4-24) were carried out in the target language.

Was your test administered on a computer? Did learners respond on computers?
The test process required continuous use of computers and technology. Students could choose to use Google Docs for their research proposal draft and Google Slides for their presentation slides. By doing so the instructor's guiding became continuous and dynamic. It was faster to provide feedback and to get into a two-way online communication. Students could also choose to use PowerPoint and Microsoft Word.

To develop their project students had to find multimodal resources, mostly online. These included social networking sites (Facebook, Instagram, YouTube, etc.), newspapers, magazines, and other webpages found with a Google search using keywords in the target language.

Was using a computer a common classroom experience for learners?
Using a computer was a common classroom experience. However, in a sense, students were also sent "into the wild" as they had to independently gather information in the target language for their project. Finding resources in the target language initially seemed an easy task for students, but soon they found it to be challenging. A workshop was provided for students on how to find multimodal resources in Italian online.

Were there any problems with the technology while learners took the test?
Students did not report issues with technology.

How did you accomplish scoring learners' performances?
Learners' spoken performance was not recorded. During the final presentation we took notes and assigned a grade based on the rubric (see Table 4-27). Here is a sample comment that we wrote during the students'

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

presentations "*The presentation slides were well structured with minor grammar errors. The student seemed a bit nervous with hesitations at times, however the pronunciation was great. The class-engaging activity Jeopardy really brought peers together and helped them review the information that was presented to them.*"

Did you score learners' spoken performances at the same time learners gave their performances?
In most cases we had enough time to score learners' spoken performances. When there was not enough time, comments were written and the scoring with the rubric was completed after the presentation.

Did you have breaks between learner performances? How did you deal with fatigue?
Learners were scheduled to present their projects at the end of the semester on specific days. All presentations had interactive activities that made the presentation less as a lecture and more enjoyable and interactive. We enjoyed the presentations. On average there were two to three presentations per day and we did not feel fatigued.

Did you ask the students themselves to score their own test? Or a classmate's test?
Students were not asked to score their own tests. However, the in-class workshops (see Table 4-24) engaged students in activities that helped them understand how they were going to be scored. For instance, learners were provided with two examples of completed presentation slides. They were asked to use the rubric, write comments, and assign a grade. Some of the comments that they wrote were: "*I don't see any class engaging activities besides a couple of questions at the beginning of the presentation*" and "*The presenter used more than 12 sources, they all seem to have come from Italian sites, and they all seem to be credible.*" From these comments we saw that learners began to critically reflect on the slides, which also helped them in the development of their own project.

Do you think your test was reliable? What did you do to check?
Since the test had detailed rubrics, we felt that our test was reliable.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

What did you do to ensure all learners had the same conditions under which they took the test?
We accommodated learners with special needs. In the case of students with presentation anxiety, the same test phases were completed as other learners, but for the final presentation the learners recorded a presentation outside of class without a live audience.

For performance tests, did you do any rater training?
There was no need to do a rater training since we were the only two instructors teaching intermediate Italian courses.

Did you make any changes to your scoring criteria as a result of colleagues' feedback during rater training?
We discussed our scoring and rubrics after the first pilot semester to improve the rubrics and make them more detailed and specific.

Reporting scores

Did you report the scores to learners? If so, how did you report test scores to learners?
Test scores were reported through the gradebook on D2L. Students were provided with feedback and points. Since the research proposal and presentation slides had a draft and a revised version, a two-way feedback took place as the learners were able to respond to comments, make changes and ask for further clarification.

What was your goal in reporting the test scores to learners?
The goal in reporting test scores to learners was to bring awareness about some aspects of their project that could be improved. This gave them the opportunity to engage in critical thinking and reflection. Figure 4-3 shows the feedback that the instructor provided to the student in the research proposal draft in Google Docs. The student then modified the draft accordingly and submitted a revised version.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

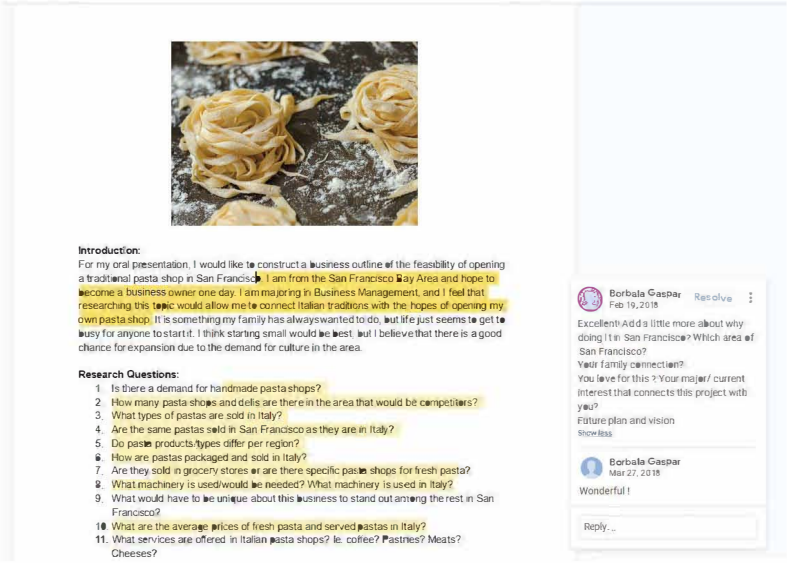


***Figure 4-3*: Screenshot showing feedback to student**

Did you spend time explaining scores, or answering learners' questions about scores in or out of class?
Yes, students had the opportunity to ask questions about their scores during office hours or via email.

How quickly did you report scores to learners? Was speed a priority?
Speed was a priority in providing feedback and scores since learners had to continue building their research project. By getting scores within ten days students felt as they had the continuous support and guidance from their instructors.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Section Three: Using Test Scores

Cut scores

How did you decide which scores were passing or failing scores (cut scores)? How did you decide which scores meant a specific grade on a test?
Failing scores would occur if students earned less than 60% in the subtests. If learners were completing the various test items in the subtests in depth, they had a high chance of receiving a high score.
Did your institution stipulate cut scores?
Yes, our institution stipulates cut scores at 60%.

For a performance test, did you use your scoring criteria to determine cut scores?
Yes, we used our scoring criteria to determine cut scores and we used the detailed rubrics to assign grades.

What was the role of the test score in determining learners' grades?
The project was 10% of the total course grade, with each subtest worth 20%. We did not decide the overall weight of the test. We simply changed the oral assessment component of the course with a multiliteracies-oriented project-based assessment while the weight stayed the same.

Using test scores

Do you feel you learned what you needed to learn from the test results about what learners could and could not do, or what they knew or did not know?
Yes, by using the rubrics and test scores it was easy to notice the areas that learners needed to work on as well as the areas in which they succeeded.

Did your test change how learners studied?
Our test produced positive washback. It promoted connections with disciplines and areas beyond the language classroom, while maintaining the course goals. In the end of the course survey for the statement "*The oral exam research project gave me opportunities to learn...*" 63% of students answered, "*a great amount of information related to the Italian culture and language.*"

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did learners ask you about the test itself (not the test scores) outside of class?
We often had one-on-one conversations about their projects. Learners mostly wanted to talk about their engaging activities, their way to synthesize their findings, and how to find more information on their topic.

Did learners' test scores change your teaching?
Since the content in each research project was specific to each student it was impossible to teach such content. Students were encouraged to attend office hours to ask questions.

Did you change the amount of class time or homework spent on specific content?
We included more in-class workshops as we felt the need to help learners make sense of the information found online in the target language.

If you could turn back time, what would you change about your test? What would you change about your test administration?
We would provide more class time for workshops. We would also include more workshops and more discussions throughout the semester. Lastly, it would be great to add more weight to the test since it is a semester-long project with numerous steps.

If you used the test and test scores for additional learning opportunities, did anything about that process help you revise the test for future use?
Yes, the test scores really showed us the areas where learners still needed help and where more time should be spent. First, in the proposal stage it would be important to spend more time on how to find and evaluate resources as well as assign learners to do additional 3-2-1 summaries of such resources (3-2-1 is asking students to describe three takeaways, two questions, and one thing they enjoyed about a specific resource). Second, learners need many practice opportunities to find information and make sense of resources. More activities should be planned to help learners get more comfortable and gain more agency and autonomy.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Section Four: Evaluating and reviewing your answers

To what extent do you think you've described recurrent patterns in your work with tests?

We think we were able to describe the recurrent patterns in our work with this test.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

An End of Chapter Quiz and a Final Examination for Beginning-Level Japanese Language Learners

Taichi Yamashita
Iowa State University

Introduction

My name is Taichi Yamashita, and I am a third-year doctoral student at Iowa State University, majoring in Applied Linguistics and Technology. My research and teaching interests focus on computer-assisted language learning and assessment in English as a Second/Foreign Language (ESL/EFL) and Japanese as a Second/Foreign Language (JSL/JFL) contexts. I have taught JSL to young learners in Japan, and to college students in the United States for two years. I am also a former EFL cram school teacher in Japan for adolescents. I am currently teaching ESL courses for international college students at Iowa State University, focusing on academic writing.

The two tests presented here were implemented in my former institution, Texas Tech University, a large public university in the United States, where I was instructor of record in first- and second-year Japanese-language courses. My students were American college students who were learning Japanese as a foreign language, with limited experience using Japanese outside of the classroom. The Japanese courses aimed to develop various aspects of communicative competence, such as expressing one's preference, and giving and receiving directions. In line with this aim, the textbook also presented some grammatical explanations and vocabulary. The classroom activities mainly involved speaking, having learners orally practice structured Japanese language in contexts which learners were thought likely to encounter in Japan.

The purpose of the end of chapter tests was to assess how well students understood a specific chapter from the textbook used in the classroom. Only one chapter test (Chapter 3 Quiz) is offered here. Particularly, the end of chapter tests were intended to measure listening comprehension, reading,

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

grammar, and vocabulary while reflecting the classroom activities. They were paper-based and included multiple choice, fill-in-the-blank, matching, and short response items. The tests were delivered after each unit was over throughout the semester. A final exam is also offered here. The final exam was intended to measure constructs which the other tests measured after each unit. Thus, the final exam consisted of test items which were very similar to what students had already been exposed to in previous tests. While there was a mid-term exam as well, limited space does not allow to offer all tests given. All the tests, including the two presented here, were used to award grades as well as to give formative feedback to students.

Because of my limited experience with testing at that time, some test items turned out to have not functioned as effectively as desired. For example, there were items which many of passing students failed to answer correctly. After rereading the test items and consulting my colleagues, I identified possible reasons for the malfunctions. Some had to do with test design, such as confusing test item order, visual prompts, and scripts, which I had not thought about when creating the test items but could have been avoided with awareness of the impact of these on test reliability. I will contribute to this collected volume by sharing my thoughts in the process of creating these test items and my reflection on these items along with potential modifications.

The Tests

The First Test: Chapter Quiz 3

1. Write a check mark in the blanks which are under a sign you think is in the museum Robert is visiting. (1X5)
   Robert and a staff in the museum are talking.

( )

( )

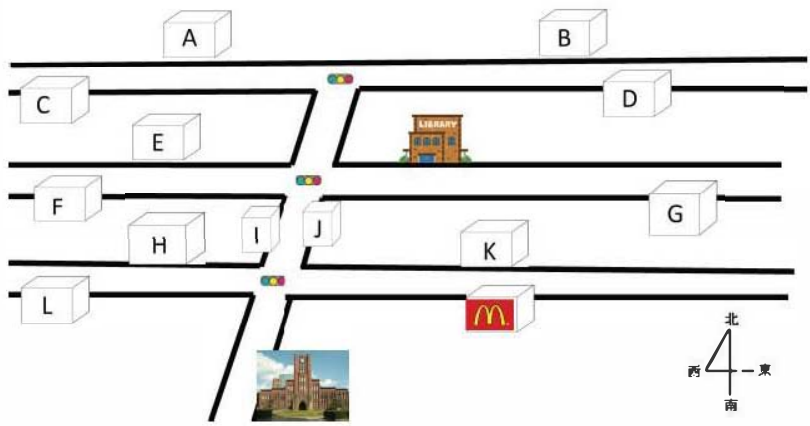
( )

( )

( )

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

2. Choose the best answer in each building based on what you hear. (1X5)
Takeshi is being asked by a stranger at a bus stop in front of University of Tokyo.



Post office ( ) Park ( ) Hotel ( ) Temple ( ) Department store ( )

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

3. Choose the best suitable particles for each blank. (1X5)

Robert and Takeshi are talking.
A: Robaato-san, tsukue _(a)_ isu o karite kite kudasai.
B: Tsukue to isu desuka.
A: Hai, tsukue o hitotsu to isu o hitotsu _(b)_ karite kite kudasai.
B: Wakarimashita. Doko desuka.
A: 207 no heya _(c)_ itte kudasai.
B: Etto, 206 janai desuka.
A: Iie, 206 _(d)_ arimasenyo.
B: Soo desuka. Jaa, 207 _(e)_ ittekimasu.

(a) □o □to □mo □wa
(b) □o □to □mo □zero-particle
(c) □de □ni □o □zero-particle
(d) □no de □no ni □no ni wa □no de wa
(e) □no de □no ni □no wa □no mo

4. Translate the words below into Japanese using hiragana. (1X2)

right away ( ) slowly ( )

5. Translate the words below into Japanese using katakana. (1X3)

personal computer ( ) shower ( )

The Second Test: Final Exam

Listening Part (35)

You will hear each dialogue twice.

I. Complete the chart below using numbers or English based on what you hear. (1X5)
Two staffs in an international student center are talking.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

	Smith, Robert	Hart, Mary	Kim, Sue
Year	Junior	Sophomore	
Age	19	21	
Major		History	Economics
Telephone number	226-0021		

II. Write prices which are suitable for each item. There are some items whose price is not mentioned in the dialogue. You must write a question mark in a blank in that case. (1X5)
Robert and a flea market clerk are talking.

(1) Watch ( )

(2) Umbrella ( )

(3) Bag ( )

(4) Jeans ( )

(5) T-shirt ( )

III. Complete the schedule for Robert below based on what you hear. (1X5)
Robert and Takeshi are talking in University of Tokyo.

A ( ) B ( ) C ( ) D ( ) E ( )

Sun	Mon	Tue	Wed	Thu	Fri	Sat
A						B
A				Today C	D	E
B						A
A						A
A						

1. Watch a Movie
2. Study Japanese
3. Read a Book
4. Play Tennis
5. Watch Tennis
6. Make a Speech
7. Listen to a Speech

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Robert and Takeshi are talking while they are listening to a speech.

Mary ( ) Prof. Sato ( ) Kay ( ) Bob ( )
Prof. Wada ( )

Speaker

	A			
B	Sue	C	D	
	E	F	G	
H	I	Ken	Prof. Suzuki	J
	Robert	Takeshi		

V. Choose an expression which describes each item the best based on what you hear. There are some items that need more than one description. (1X5)
Robert and Takeshi are talking about Robert's trip to Kyoto.
a. Room □new □clean □old □big
b. Dinner □delicious □expensive □not delicious □not expensive
c. Weather □hot □cold □good □not good
d. Kabuki □good-looking □interesting □fun □boring
e. Temple □beautiful □lively □quiet □old

VI. Write a check mark in each blank which is under a sign which you think is in the museum. (1X5)
Robert and a staff in a museum are talking.


( )


( )

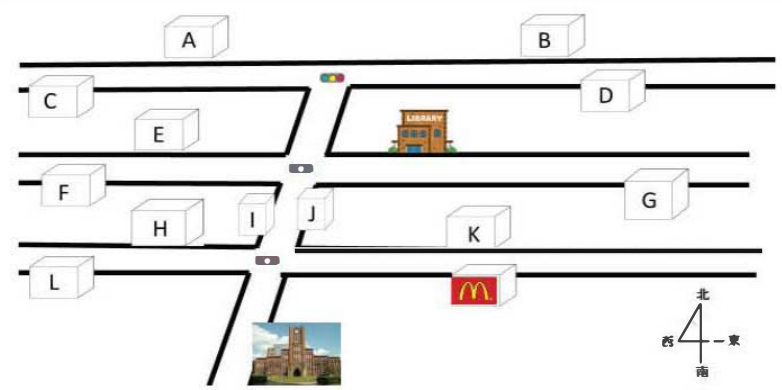

( )

( )

( )

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

VII. Write a letter which represents the location of each building the best. (1X5)
Robert is asked by a stranger at a bus stop in front of University of Tokyo.
Post Office ( ) Bookstore ( ) Temple ( ) Restaurant ( ) Hospital ( )



Reading Part (10)

I. Read the dialogue below and answer each question. You can use English in answering. (1X5)
Robert and Takeshi are talking in University of Tokyo.

A: あ、ロバートさん、こんにちは。
B: こんにちは、たけしさん。
A: あ、デパートに行きませんか。
B: んーちょっと。サクラホールで北山先生のスピーチがありますから。
A: 今日は西川先生のじゃないですか。
B: いいえ、北山先生のですよ。
A: そうですか。あの、私も行っていいですか。
B: いいですよ。じゃあ、カフェテリアの中にいてください。すぐ行きますから。
A: わかりました。じゃあ、カフェテリアにいますね。

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

1. What does Robert invite Takeshi to do?

   _____.

2. Why does Takeshi decline Robert's invitation?

   _____.

3. Who is giving a speech today?

   _____.

4. What does Takeshi say in line 7?

   _____.

5. Where will they meet after a while?

   _____.

II. Read the dialogue below and choose the most suitable particles on each blank. (1X5)
Robert and Takeshi are talking in University of Tokyo.

A: Robaato-san, kyoo Suzuki-sensee no supiichi ikimasuka.
B: Watashi wa Sato-sensee _(a)_ ikimasu. Dakara Suzuki-sensee _(b)_ ikimasen nee.
A: Soo desuka. A, 207 no heya _(c)_ tsukue hitotsu _(d)_ isu futatsu _(e)_ karite kite kudasai. Genki kurabu ga arimasu kara.
B: Wakarimashita.

(a) □ni □o □no ni □no o
(b) □no ni □no o□ no ni wa □no ni mo
(c) □ni □de □o □e
(d) □o □mo □to □no o
(e) □o □mo □to □no o

Writing Part (15)

I. Translate the words below into Japanese using katakana. (1X5)

1. e-mail ( )
2. restaurant ( )
3. bus ( )

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

4. part-time job (　　)

5. shower (　　)

II. Translate the words below into Japanese using hiragana. (1X5)

1. extremely (　　)

2. and then (　　)

3. right away (　　)

4. slowly (　　)

5. many (　　)

III. Write the appropriate mixes of kanji and hiragana (if necessary) for words on each underline. (1X5)

1. これは 980 えん(kyuuhyaku hachijuu en)です。
(　　)

2. あのクラスはすいようび(suiyoobi)の 5 じはん(gojihan)です。
(　　/　　)

3. がいこく(gaikoku)のひと(hito)がたくさんいますね。
(　　/　　)

4. パスタをたべて(tabete)、ワインをのみました(nomimashita)。
(　　/　　)

5. ひがしぐち(higashiguchi)をでて(dete)、ひだり(hidari)に 3 ぷん(sanpun)です。
(　　/　　/　　/　　)

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

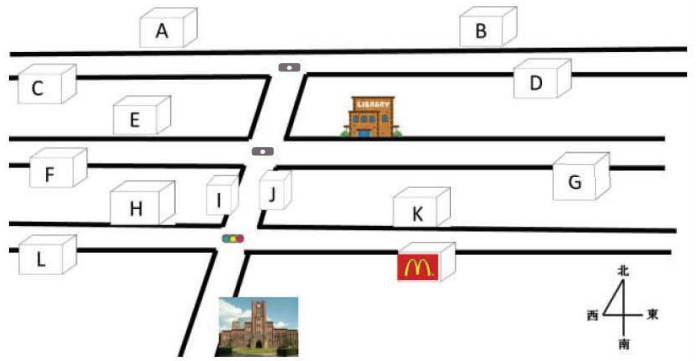
English Translation and Answer Key of the First Test (Chapter Quiz 3)

1. Write a check mark in the blanks which are under a sign you think is in the museum Robert is visiting. (1X5)

Robert and a staff in the museum are talking.

2. Choose the best answer in each building based on what you hear. (1X5)

Takeshi is being asked by a stranger at a bus stop in front of University of Tokyo.



Post office (*K*) Park (*J*) Hotel (*E*) Temple (*D*) Department store (*I*)

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

3. Choose the best suitable particles for each blank. (1X5)

Robert and Takeshi are talking.
A: Robaato-san, tsukue _(a)_ isu o karite kite kudasai.
B: Tsukue to isu desuka.
A: Hai, tsukue o hitotsu to isu o hitotsu _(b)_ karite kite kudasai.
B: Wakarimashita. Doko desuka.
A: 207 no heya _(c)_ itte kudasai.
B: Etto, 206 janai desuka.
A: Iie, 206 _(d)_ arimasenyo.
B: Soo desuka. Jaa, 207 _(e)_ ittekimasu.

A: Robert, please go borrow a table and chair.
B: A table and chair?
A: Yes, please go borrow one table and one chair.
B: I understood. Where?
A: Please go to Room 207.
B: Um... isn't it Room 206?
A: No, it's not 206.
B: I see. Then, I'm going to 207.

(a) □o □*to* □mo □wa
(b) □o □to □mo □*zero-particle*
(c) □de □*ni* □o □zero-particle
(d) □no de □no ni □*no ni wa* □no de wa
(e) □no de □*no ni* □no wa □no mo

4. Translate the words below into Japanese using hiragana. (1X2)

right away (すぐに) slowly (ゆっくり)

5. Translate the words below into Japanese using katakana. (1X3)

personal computer (パソコン) shower (シャワー)

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

English Translation and Answer Key of the Second Test (Final Exam)

Listening Part (35)

You will hear each dialogue twice.

1. Complete the chart below using numbers or English based on what you hear. (1X5)

Two staffs in an international student center are talking.

	Smith, Robert	Hart, Mary	Kim, Sue
Year	Junior	Sophomore	*Senior*
Age	19	21	*22*
Major	*Business*	History	Economics
Telephone number	226-0021	*301-7809*	*547-1269*

2. Write prices which are suitable for each item. There are some items whose price is not mentioned in the dialogue. You must write a question mark in a blank in that case. (1X5)

Robert and a flea market clerk are talking.

(1) Watch (*600*)

(2) Umbrella (*blank*)

(3) Bag (*4500*)

(4) Jeans (*blank*)

(5) T-shirt (*1200*)

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

3. Complete the schedule for Robert below based on what you hear. (1X5)
Robert and Takeshi are talking at the University of Tokyo.

A (*2*) B (*4*) C (*5*) D (*6*) E (*1*)

Sun	Mon	Tue	Wed	Thu	Fri	Sat
A						B
A				Today C	D	E
B						A
A						A
A						

1. Watch a Movie
2. Study Japanese
3. Read a Book
4. Play Tennis
5. Watch Tennis
6. Make a Speech
7. Listen to a Speech

4. Write a letter which represents a suitable location where each person is. (1X5)

Robert and Takeshi are talking while they are listening to a speech.

Mary (*F*) Prof. Sato (*J*) Kay (*C*) Bob (*E*)
Prof. Wada (*H*)

Speaker

	A			
B	Sue	C	D	
	E	F	G	
H	I	Ken	Prof. Suzuki	J
	Robert	Takeshi		

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

5. Choose an expression which describes each item the best based on what you hear. There are some items that need more than one description. (1X5)

Robert and Takeshi are talking about Robert's trip to Kyoto.

f. Room □new □*clean* □old □big
g. Dinner □*delicious* □expensive □not delicious □not expensive
h. Weather □hot □cold □*good* □not good
i. Kabuki □good-looking □interesting □fun □*boring*
j. Temple □*beautiful* □*lively* □quiet □old

6. Write a check mark in each blank which is under a sign which you think is in the museum. (1X5)

Robert and a staff in a museum are talking.


( )

( )

( )

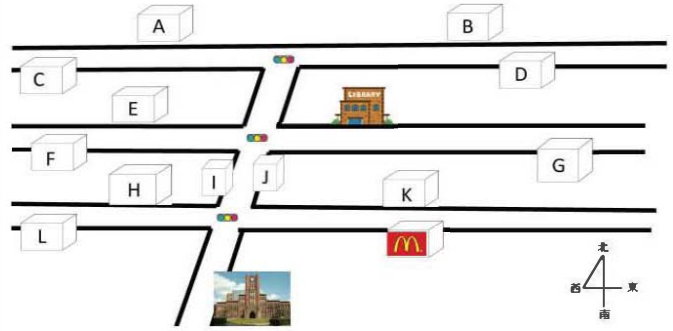
( )

7. Write a letter which represents the location of each building the best. (1X5)

Robert is asked by a stranger at a bus stop in front of University of Tokyo.

Post Office (*I*) Bookstore (*F*) Temple (*H*) Restaurant (*A*) Hospital (*B*)

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use



Reading Part (10)

I. Read the dialogue below and answer each question. You can use English in answering. (1X5)

Robert and Takeshi are talking in University of Tokyo.

A: あ、ロバートさん、こんにちは。
B: こんにちは、たけしさん。
A: あ、デパートに行きませんか。
B: んーちょっと。サクラホールで北山先生のスピーチがありますから。
A: 今日は西川先生のじゃないですか。
B: いいえ、北山先生のですよ。
A: そうですか。あの、私も行っていいですか。
B: いいですよ。じゃあ、カフェテリアの中にいてください。すぐ行きますから。
A: わかりました。じゃあ、カフェテリアにいますね。

A: Hi, Robert, Good afternoon.
B: Hi Takeshi.
A: Why don't we go to a department store?
B: Sorry, I can't. There is Prof. Kitayama's speech in Sakura Hall.
A: Isn't it Prof. Nishikawa's speech today?
B: No, it's Prof. Kitayama's.
A: Is that so? Can I go with you?
B: Okay. Then, please be in the cafeteria. I'll go there soon.
A: I see. I'll be there.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

6. What does Robert invite Takeshi to do?
To go to a department store.

7. Why does Takeshi decline Robert's invitation?
Because he is going to listen to a speech.

8. Who is giving a speech today?
Prof. Kitayama.

9. What does Takeshi say in line 7?
He asks if he can go with Robert.

10. Where will they meet after a while?
In the cafeteria.

II. Read the dialogue below and choose the most suitable particles on each blank. (1X5)
Robert and Takeshi are talking in University of Tokyo.

A: Robaato-san, kyoo Suzuki-sensee no supiichi ikimasuka.
B: Watashi wa Sato-sensee _(a)_ ikimasu. Dakara Suzuki-sensee _(b)_ ikimasen nee.
A: Soo desuka. A, 207 no heya _(c)_ tsukue hitotsu _(d)_ isu futatsu _(e)_ karite kite kudasai. Genki kurabu ga arimasu kara.
B: Wakarimashita.

A: Robert, do you go to Prof. Suzuki's speech today?
B: I go to Prof. Sato's. So, I don't go to Prof. Suzuki's.
A: I see. Please borrow one table and two chairs for 207 because we have Genki club today.
B: I understood.

(f)	□ni	□o	□*no ni*	□no o
(g)	□no ni	□no o	□ *noni wa*	□no ni mo
(h)	□ni	□*de*	□o	□e
(i)	□o	□mo	□*to*	□no o
(j)	□*o*	□mo	□to	□no o

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Writing Part (15)

I. Translate the words below into Japanese using katakana. (1X5)

6. e-mail (メール)

7. restaurant (レストラン)

8. bus (バス)

9. part-time job (アルバイト)

10. shower (シャワー)

II. Translate the words below into Japanese using hiragana. (1X5)

6. extremely (すごく)

7. and then (それから)

8. right away (すぐ)

9. slowly (ゆっくり)

10. many (たくさん)

III. Write the appropriate mixes of kanji and hiragana (if necessary) for words on each underline. (1X5)

6. これは 980 えん(kyuuhyaku hachijuu en)です。
(九百八十円)

7. あのクラスはすいようび(suiyoobi)の 5 じはん (gojihan)です。
(水曜日 / 五時半)

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

8. がいこく(gaikoku)のひと(hito)がたくさんいますね。
(外国 / 人)

9. パスタをたべて(tabete)、ワインをのみました(nomimashita)。
(食べて / 飲みました)

10. ひがしぐち(higashiguchi)をでて(dete)、ひだり(hidari)に3ぷん(sanpun)です。
(東口 / 出て / 左 / 三分)

1. This is 980 yen.
2. That class is at 5pm on Wednesdays.
3. There are many foreigners.
4. I had pasta and drank wine.
5. You should get out from the east exit and walk for three minutes.

Contributor's Questionnaire Responses

Section One: Test Planning and Writing

Why did you write the test? What were the purposes of the test?
Consider: To help learners focus their thoughts, or change how they studied?
To award a course grade, or part of a grade?
To learn whether learners met a course objective?
To use the scores to give learners feedback on their progress?
Something else?
These tests were created primarily to assess learners' proficiency and thereby to award a grade. At the same time, however, the tests also aimed to have students check to what extent they achieved the expected learning outcomes of a certain unit in the textbook used for the course. For this sake, students received feedback on their tests during the class hours, although that was not the primary purpose.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

How did you decide how many subtests to write? Was there some correspondence between a subtest and a course objective, or was there some other reason to create the subtests you did?

The chapter quizzes usually consisted of four subtests. The first two subtests were intended for assessing students' listening comprehension. The second subtest involved reading comprehension, and the last two subtests were supposed to assess students' vocabulary knowledge. These subtests were created given the specific course objectives shared with students in the syllabus. For example, the syllabus explicitly listed the listening comprehension ability as one expected outcome of this course. Also, the syllabus described that students will be able to "comprehend context of some reading material by using Hiragana, Katakana, and Kanji which they learned." Regarding vocabulary, students were expected to learn 300 words along with all the Japanese characters of Hiragana and Katakana and 58 Kanji characters. Given these learning outcomes listed in the syllabus, the chapter quizzes were intended to assess what students were expected to achieve for each of these specific outcomes listed in the syllabus.

Regarding the final exam, it was intended to tap into course content but in a more comprehensive manner. For this sake, the final exam consisted of each component of chapter quizzes in the semester. For example, the last two listening comprehension items in the final exam were adopted from one of the chapter quizzes (see Chapter 3 Quiz), though the dialogue to be heard was different. Thus, it was expected that students were familiar with the format but still required to work on new dialogues or audio stimuli. The final exam did not include items which were completely new to students.

Did you write as many items or test tasks as you needed, or did you write more than you needed, then pare the number down? How did you decide which items or test tasks to keep? How did you decide which items or test tasks to discard?

I did not start with writing as many items as I needed. Rather, I firstly decided on the number of items for one single chapter quiz or test. This decision was made based on class hours available for the test implementation. Then, because the course emphasized oral tasks including listening and speaking, I decided to write more items on listening than items on reading and vocabulary. At the point of the implementation, I usually spent 20-30 minutes to administer each test. Most of the students did not use the whole 30 minutes, but some students fully used the 30 minutes. Regarding the final exam students had 50 minutes.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

How did you decide how many items to write for each subtest?
Given the time constraints of 20-30 minutes, I decided to write 20 items in total for each chapter quiz. Then, since regular class hours were mostly spent on listening and speaking, I decided to write 10 items for listening comprehension subtests and five items for each reading and vocabulary subtests so that the weighted importance of listening in the classroom is reflected in the test at least.

For the number of each subtest in the final exam, I did not have a strong rationale. Rather, given 50 minutes allowed to administer it, I intuitively thought that about five items per subtest would be reasonable. Despite this intuitive decision making, I did not encounter significant issues.

How did you decide how many test items to write in total?
As mentioned above, I based the number of items on the class or exam hours available for the test. Because most of my students had received no formal instructions on Japanese in the past, I expected them to spend more time.

Did you have one version of your test, or did you create a second equivalent version?
I did not create a second equivalent version.

Were you concerned at how long the test would take to administer?
Yes. This is especially because I was not an experienced teacher back then and had to create my own test, instead of using an off-the-shelf test. As I got more experience in the instutition, however, I gradually became confident in estimating the amount of time learners would spend.

Were you concerned at how long the test would take to score?
No. Since my students were beginners, their responses to the test were limited and thus easy to score for me as a L1 speaker of the target language.

Were you concerned how you might use the test items themselves for learner feedback?
I think I was. However, I was more concerned about giving a grade than giving feedback. Although I spent class hours on explaining about each item after returning the test to the students, I spent much more time on thinking about how to match what I was doing during class hours and what the students were to be tested on. This way of thinking naturally led me to consider the test to be an assessment tool rather than a learning tool.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you consider having your students take your test on a computer? Why or why not? If not, what were your concerns?
No, I did not consider this possibility. Firstly, I was not used to a computer itself for testing purposes. So, even if I had had this idea, I would not have implemented a test on a computer because of potential unanticipated problems, such as data loss and equipment malfunction. Secondly, I did not know how to create computer-delivered tests. The only way I could have created computer-delivered tests was to use PowerPoint, which I do not think is sophisticated enough. Thus, given my concerns about data and limited resources I could handle, I did not consider computer-delivered tests.

Did you consider making the test an open book test?
No, I didn't. This is because I thought that students should use their own knowledge during the test.

Did you consider allowing learners to use additional sources such as dictionaries, or their notes, while taking the test?
No, I didn't for the same reason as the previous response.

Did you plan to allow learners to re-take a test for improvement? The same test, or a different test?
No. I did not have time to implement a second test in my class hours because I needed to move forward to finish the course content. Also, I do not think that I had such an idea first of all.

What sources did you draw from for your test items?
Consider: Test item ideas or content from previous tests?
Test item ideas or content from a textbook?
Ideas or content from review sections of a textbook?
Test item ideas or content based on textbook activities?
Test item ideas or content from course objectives?
Test item ideas or content from what learners do in class?
I created my test primarily based on classroom activities. Some of them were adopted from the textbook, but others were my own activities. I tried to make a good connection between what the students were doing in their class hours and what they were expected to do in the tests. So, I think my response to the above questions should be positive, even though I was more concerned about the relationship between what learners were expected to perform in the classroom and what they were expected to do on a test.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Test item ideas or content from what learners do for homework?
Whereas there may have been similarities between what students were working on as homework and what they were doing in the tests, I was not aware of similarity if any. Regarding hiragana, katakana, and kanji writing, because I could spend much class hours on this aspect, students' practice on this was primarily done as homework. In that sense, the writing subtest for the final exam was adopted from the homework learners were working on.

Test item ideas or content from worksheets you make?
Since I did not create many worksheets for my teaching, this question may not be applicable to my case. Still, I was trying to connect what students were expected to perform in the class hours with what they were expected to perform in the tests. For instance, I created one of the visual prompts on a test based on a PowerPoint slide I used for a lesson.

Did you consider learners' communicative competence when writing test items?
For the listening items, I think I was thinking about the communicative competence. I tried to create test items in the way that better performance on test items indicate learners' communicative achievement expected in the real context (e.g., identifying a location of items on a map).

What aspects of communicative competence?
The listening comprehension asked learners to do something with audio prompts. For instance, in the Chapter 3 Quiz, students were asked to identify the location of buildings based on a dialogue between two people. Chapter 3 Quiz also had students choose signs that indicated what they must not do in a certain place. Although the items themselves do not measure learners' language production, it would be reasonable to claim that the test items measure communicative competence because learners are likely to encounter this kind of context in the real life (e.g., identify a landmark on a map, identify prohibited activities).

Which test item formats do you prefer to use?
Consider: Short answer, fill-in-the-blank, matching, cloze, performance test, etc.?
I think my preference may depend on what constructs I am trying to measure. For instance, when there are several possible answers for one item, I would create a test item in the form of short answer. In contrast, if there is only one single answer for the item, I am likely to create the test item in the form of matching.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

However, when I consider the expected rating workload, I prefer not to create short answer items because students' responses are not always predictable, and thus I would be required to distinguish responses which deserve the full credit and those which do not but deserve the partial credits.

What types of learner knowledge do you believe you are capturing in your test?
I had three major constructs in my mind when I created these tests. First, the tests were intended to measure listening comprehension. All the tests shared start with listening comprehension items where students were told to respond to items based on audio stimuli. Secondly, the tests, particularly later chapter quizzes and the final exam, were supposed to measure reading comprehension, requiring learners to understand not only linguistic features but a passage as a whole. The third knowledge to be measured was the ability to write Japanese translations for English words using Japanese characters. This component was aimed to measure vocabulary knowledge and spelling knowledge.

Did you seek help from a peer to clarify what your test items or tasks were measuring?
Consider: Did you state at any point what you wanted to measure in your test?
I created the tests without consulting my colleagues. Also, I do not think they asked me for help with their tests except for recording a dialogue for listening comprehension items.

Did you make any changes to your test or items as a result of your colleague's feedback?
I did not make any changes on my test before the implementation. However, after implementing the tests, I consulted my colleague. Although I received feedback from them, I did not have a chance to revise or reimplement the test thereafter.

The colleague suggested that the reading text on a mid-term exam was considerably long, and the information was dense. The same feedback was given on one of the listening comprehension items on the final exam. If I had had another opportunity to implement the test, I would have modified the test items based on his feedback. I did not have a chance to get feedback on chapter quizzes. In fact, before I started to analyze my tests, I was uncertain about how to do that. So, with this lack of knowledge and time constraints from regular teaching duties, I did not have ample opportunities to get feedback on chapter quizzes.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you compare your test to the lessons that learners had?
Yes. I often made the comparisons of this sort while I was creating the tests in an attempt to ensure the relationship between classroom activities and test items.

Did you compare your test to the textbook or other materials learners used?
Yes. As indicated in the previous question, I frequently thought about the connection between the classroom instructions and the test items. However, rather than referring to the textbook, I think I relied more on materials I created and implemented in my class hours.

Section Two: Test Administering and Scoring

Were you concerned about test security? What did you do to ensure test security?
I was not concerned about test security. I think students had no access to the tests before its implementation. I simply stored the test data in a USB memory stick.

Were your tests photocopied, or did learners see your test items or test tasks another way, such as on the blackboard?
I did not give my students any access to the tests before the implementation. However, after I returned their tests, I shared the tests and audio materials on Dropbox.

How did you prepare learners to take the test? Were there any test items or test procedures that were new to them?
For the very first test, I think I had them take a practice test to familiarize them with the test format.

Did you pilot your test? Do a trial run? Did the pilot result in any changes to the final version of your test?
I did not pilot my test. However, as I indicated in the previous question, I exposed students to a practice test. In this sense, I piloted the practice test. I do not think I found issues with the test format. So, I followed the format thereafter.

Did you write any of the test in the learners' first language? Why?
Yes. All the instructions were written in English. English was not the first language for all the students because some of them spoke languages other than English as their first language, though English was the language

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

commonly spoken by all the students. This was because I did not want to measure the ability to understand the test instructions. Also, it was almost impossible to write instructions using Japanese language that they had already learned. Rather, I wanted to measure the language ability they have developed in the class hours. Their target language was used only in the reading comprehension in the final exam. This was intentionally done, and the students were informed in advance that the reading text would be completely written in Japanese. The ability to read a Japanese discourse was part of the construct.

For classroom tests: How did you accomplish scoring learners' tests? Consider: Did you score learners' tests twice for accuracy?
No. The scoring was done only once. Since the scoring was relatively straightforward, I never thought about this idea.

Did you write a test key?
I think so. But I am afraid that I do not keep the data now.

Did you consider alternate answers and add them to the key?
I had to think about alternate answers every time I encountered a response that I had not expected. This was especially the case for short answer items. Although I thought that possible answers were considerably limited, I encountered a range of responses.

Did you hide learners' names as you scored?
No. I scored tests with being aware of whose test they were.

Did you randomize learners' tests for scoring, thus erasing any order from where learners sat, or at what point they handed their test in?
No. I simply scored tests in the order of submission (or backwards).

Did you go back and change your marks on previously scored tests in response to problems you found while scoring tests later in the process?
As far as I remember, I tried not to go back and change marks and tried to stick to the criteria I had created at the point of creating the tests. This was because I did not want to change the criteria frequently given that I spent much time on creating the criteria. However, I guess that there were a few cases where I modified my mark, especially for short answer items, given some unexpected but reasonable responses from my students.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you later change your scoring because you recognized some learners couldn't answer certain test items due to some aspect of the items or the test?
As far as I remember, I do not think I changed my scoring for the reason of this sort.

Did you put learners' responses to items into a spreadsheet for further analysis? Did that process help you catch scoring accuracy problems? Or problems with bias?
When I scored the tests, I did not use a spreadsheet at any point. When I analyzed my test items in a graduate class, I wished I could have done this.

Did you ask the students themselves to score their own test? Or a classmate's test?
I asked students to evaluate themselves for an oral proficiency test (not presented here), but I did not do so for the paper and ink format tests that I present here.

Do you think your test was reliable? What did you do to check?
Consider: Did you ask another teacher to look over your test before you administered it? What did he or she say? Did you make any changes to your test as a result?
What did you do to ensure all learners had the same conditions under which they took the test?
For a classroom test, did you use item analyses on a spreadsheet such as item facility, or B-index?
After implementing the tests, I analyzed the test items using item facility and B-index. This process helped me identify potentially problematic test items, whose B-index is negative, and a potential underlying reason. For instance, I found that sometimes the discourse texts for the listening comprehension may have been too long for students to follow easily.

For a classroom test, did you use any statistical reliability estimates, such as Cronbach's alpha?
I used phi(lambda) dependability.

Reporting scores

Did you report the scores to learners? If so, how did you report test scores to learners?
Yes. I wrote a score on each test sheet and got the sheet back to students.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Reporting scores

What was your goal in reporting the test scores to learners?
I wanted them to use the score as part of feedback. Because I reported the average score, students could know how well they performed in relation to other classmates.

Did you teach learners how to interpret their test scores?
I did not teach them how to interpret their test scores. However, I reported the average score so that they would get a better sense of how well they performed relative to others.

Did you report the scores to anyone else?
No, I did not report scores to anyone else.

Did you spend time explaining scores, or answering learners' questions about scores in or out of class?
I went through all the items on the tests with students in class hours. I did not explain the scores, but I think that explaining about all the test items may have clarified the scores at the same time.

How quickly did you report scores to learners? Was speed a priority?
I did not time the scoring, but I do not think that I spent more than 30 minutes for scoring about 15 papers for the chapter quiz, which included 20 items. The speed was not priority. Basically, I finished my scoring on the day when the test was implemented.

Section Three: Using Test Scores

Cut scores

How did you decide which scores were passing or failing scores (cut scores)? How did you decide which scores meant a specific grade on a test? Consider: Did a language use framework such as CEFR or other standards help you determine cut scores?
Did your institution stipulate cut scores?
I did not follow any external sources to decide on the cut scores. Rather, I primarily relied on my intuition that students who scored higher than 60% would be able to keep up with the subsequent chapters and to achieve some communications in Japanese language which they had learned by that point of time.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

How did you use learners' scores from this test?
Consider: Were the scores for your use only?
I reported the scores to the students with the average score so that they would be able to see how well they performed in relation to other classmates. I hoped that this report would be part of the formative assessment for themselves and thus that they would realize how much effort they were expected to put thereafter.

What was the role of the test score in determining learners' grades?
Consider: How much weight did you give your test? How did you decide?
The chapter quizzes taken together accounted for 20% of the final grade. Because of my limited teaching experience at the institution back then, I simply referred to a past syllabus to make the decision. Meanwhile, the final exam accounted for 20% of the final grade. Like the chapter quizzes, I made this decision based on the past syllabus I consulted.

Were other measures used to decide learners' grades, besides your test?
Yes. The final grade consisted of classroom participation, homework, quizzes, mid-term exam, and final exam with each accounting for 20%. The tests that are shared in this volume is one of the six chapter quizzes and the final exam.

What was the relationship of the other measures to your test?
The classroom participation is mostly related to the listening comprehension given that the classroom instruction primarily involved oral interactions. The homework mostly included practice on Japanese writing system (i.e., hiragana, katakana, kanji), so this has to do with the translation part of the tests. However, the homework was not used to measure learners' ability. Rather, it was implemented to make sure that learners practiced by themselves before the class starts.

Did your test capture some knowledge, skill, or ability the other measures did not capture?
For the classroom participation, it was often the case that students relied on their classmates when they could not understand what I said or when they could not respond to my questions in Japanese. Meanwhile, the tests measured their independent ability to perform something in Japanese because they could not consult their textbook or classmates. Therefore, the tests captured their independent performance which was not measured by other classroom assessments.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Reporting scores

How did you report scores to learners? Was timeliness of concern to you?
I marked their score on the corner of each paper and got them back to the students in a way that other classmates could not see other students' score. The mark on the paper was the only way for them to know the score. Timeliness was not such a concern, but I was trying to get the tests back to them as soon as possible. In fact, a test was usually implemented on Friday and returned to the students on Monday.

Did you hand the test back to learners? Did the learners get to keep the tests? Or did you take the tests back?
Yes. I handed the test back to each student in the class hours. I did not tell them to keep the tests for their record. It was totally up to them. Once I handed the tests back to the students, I did not get the tests back again.

Did you offer feedback to individual learners in addition to their test scores? Written? Orally? In or out of class?
I simply wrote the answer for each item as feedback. In the class hours, I orally explained each item so that the students could ask questions about them in the class hours.

Did you teach learners to interpret their scores?
I did not. The only thing I did for them to interpret the score was that I showed the average score so that students could know how well they performed relative to their peers.

Using test scores

Do you feel you learned what you needed to learn from the test results about what learners could and could not do, or what they knew or did not know?
I think so, but this feeling may be very intuitive. That is, I could predict how well each student would perform on a test based on their classroom performances.

Did your test change how learners studied?
Consider: Did you use learners' scores to find out if your test caused washback?
I do not think so, especially because I prioritized the connection between the classroom instruction and test items. That is, in order for them to get high scores on the tests, they should have performed well in the classroom

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

activities throughout the semester. From my perspective, if they performed well in the class hours, they should have scored higher in a test.

Did you use particular item types or a performance test to change learners' practices or support their learning? Did their scores indicate they had changed their learning practices?
When I was creating test items, I did not explicitly aim to change their practices or support their learning, given that the class was originally designed in the way that promotes oral language practice. I do not think that changes in their scores indicated their learning practices changed.

Did you spend time going over the test in class?
Consider: Did you mention trouble points as general comments?
Yes. The class hours were spent on going over each item. I played listening passages as well so that students could confirm the answers. I think I mentioned some test items that many students got wrong.

Did you go over each item?
Yes.

Did you go over specific subtests?
I went over all the items. However, I did not review the section where students were asked to write Japanese translations for each English word.

Was going over the test a classroom activity?
No. I conducted the session mostly in a teacher-fronted way.

Did learners ask you about the test itself (not the test scores) outside of class? If so, what did they want to talk to you about? Did you use specific subtests or items or tasks to focus your talks with learners?
As far as I remember, they did not ask me about the test itself outside of class.

Did learners' test scores change your teaching?
Consider: Did you re-teach content because learners didn't do well on your test?
I do not think so. The review on the test items was simply going over the test items without a focus on particular items.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you skip content because learners did well on your test?
I do not think so. I covered all the test items (but the writing section) regardless of students' performance on them.

Did you change the amount of class time or homework spent on specific content?
No. I did not distribute the amount of class hours in that way.

Did you change your teaching for future courses based on test results?
I may have changed my teaching, but I am not sure if this resulted from the test results.

If you could turn back time, what would you change about your test? What would you change about your test administration?
Consider: Did learners give you feedback on the test? Did they think the test was fair, or helpful?
I did not hear much of students' voice about the test items, so my modification of the test items would be based on post-hoc analyses I did as part of my master program. For instance, I found that the order of test items in listening comprehension printed on the test was not in line with the order of their answers cued in the audio materials, which may have resulted in the fact that passing students did not get these items right.

Did others (parents or administrators or colleagues) give you feedback on the test?
One of my colleagues said that my listening comprehension dialogues and reading comprehension texts were very long for students. I appreciated his feedback, so if I have another chance to teach Japanese, I will definitely incorporate his feedback.

If you used the test and test scores for additional learning opportunities, did anything about that process help you revise the test for future use?
Yes. As I suggested in the last question, I took a second language testing course in the master program. In that class, I analyzed the tests shared in this collected volume, using some statistical analyses I had not known before. Through this process, I identified some issues with test items as indicated in the previous item.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Section Four: Evaluating and Reviewing your Answers

Please read through your answers to the items and answer the following:
To what extent do you think you've described recurrent patterns in your work with tests?
From what I can see above, it seems that there are two aspects that repeatedly emerge. First, I think I tried to make sure that what students were expected to perform on a test is strongly related to what they were expected to perform in classroom hours. Secondly, I benefited from asking my colleagues for their comments on my test items, especially because I had not had such opportunities before.

To what extent is your test here an innovation, or something new, for you?
I do not see much innovation in the tests I shared. I would say that the tests are somewhat traditional but may be different from other tests commonly used for L2 Japanese learners given its focus on listening comprehension. I do not know much about tests for L2 learners of Japanese, so it would be worthwhile looking into them.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

A Written and Oral Russian Achievement Test for Beginning College-level Learners

Irina Drigalenko
Texas Tech University

Introduction

I am a full-time Russian instructor at a large public university in the southwestern United States. I also teach intensive summer courses at a Critical Languages Institute in a nearby state. I am a native speaker of Russian with many years of experience of teaching beginning and intermediate Russian college courses. I am passionate about teaching Russian, and my research focuses on students' motivation for studying the Russian language. I hold two Master's Degrees. The first is from Tomsk State University, Russia in Pedagogy and Philology and Russian Language and Literature, and the second is from Texas Tech University in Applied Linguistics. Teaching Russian Language as a foreign language became my passion from day one.

When I started teaching as a part-time instructor, I used inherited tests, but always modified them to the level at which my students were able to perform. Even now, I prefer to design or redesign existing tests for my students based on their level of proficiency. Each group of students I teach has specific needs, levels of motivation, and paces of acquisition. Designing and redesigning the tests is, of course, a challenge, but it is also fun. I try to design tests to fit the way my students might experience the need to apply material for real life application– challenging, but useful, where test activities connect them to the purpose of learning the language. It makes us co-creators of the language realm.

The Russian Program at our university provides students with knowledge of Russian, from Elementary courses to Advanced courses. The Beginning second semester of Russian 1502 meets five times a week (50-minute periods.) Class size varies each semester, but we usually have up to

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

20 students in each class. For preparation of this represented test, we used the textbook *Troika: A Communicative Approach to Russian Language, Life, and Culture* (2nd ed.) (Nummikoski, 2011). This textbook is suitable for first-year students since it provides a very good structure to the learning process. Students who took the Chapter 11 Test presented here[1] were 23 college students age 18-25. They chose and continued with Russian for various reasons (for future Homeland Security careers or ROTC scholarship opportunities, to challenge themselves with a less commonly taught language, to keep up with the heritage of their family).

The chapter tests we use in our Program provide useful feedback for our students. They help with learners' motivation and focus. They also help me, as an instructor, to monitor my learners' progress and teach more effectively. I wrote the exam presented here for the students of a beginning Russian course, to be taken after learning the material based in Chapter 11 of *Troika.* In this chapter, students learned and were able to use the names of various foods and beverages in Russian that go along with the verbs for "to eat," "to drink," "to buy," and "to sell." They also learned how to count money, compare prices, and shop for food. Students practiced broader meanings of the genitive case to quantify objects, to denote the absence of something, or to express negation. Additionally, students learned how to better use the accusative case versus the nominative case depending on the meaning of a communicative task, and ways to modify adjectives in order to compare objects. The test was comprised of five sections: (I) Oral presentation, (II) Grammar, (III) Size and Cost interpretation, (IV) Essay/Dialogue, and (V) Bonus.

The Test

RUSN 1502 Экзамен Урок 11 Spring 2018

Имя, фамилия _____.

_____**/100**

I. Oral presentation _____**/20 pts.** Communicative task: *You and a friend (or any family member) are discussing an upcoming party on the phone and deciding what you need to buy, how much you need, and who is buying what.*

[1] Thanks to Matthew Horn for his timely assistance.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

II. Грамматика. Language control.

1._____/15 pts. **Что вы обычно едите и пьёте на завтрак, на обед и на ужин?** *What do you usually eat and drink for breakfast, lunch, and supper?*
List 3 food items and 1 drink item for each meal. If you never eat your breakfast, add food items to your other meals. Answer in full sentences. Use verbs **есть** *и* **пить.**

На завтрак  (3)

_____

_____

 (1)_____

На обед  (3)

_____

_____

 (1)_____

На ужин  (3)

_____

 (1)_____

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

2._____/10 pts. Поговорнм о еде. *Let's talk about food. Answer following questions in full sentences:*

1.Чего вы не любите есть?

_____

_____

2 .Какие напитки вы обычно пьёте?

_____

_____

3. Какие фрукты или овощи вы едите каждый день?

_____

_____

4.Что вы ели вчера на обед?

_____

_____

5. Какую еду можно купить в кулинарии?

_____

_____

3. _____/ 5pts. *Match containers to the items. Use the proper case for the quantity of containers.*

пакет	бутылка	коробка	банка	пачка

Вчера мы купили_____ красного вина,

_____ молока,

_____ шоколадных конфет,

_____черной икры

и три маленьких _____чёрного чая.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

4. _____/ 20 pts. *Provide appropriate case forms for the following food items:*

Food / Еда	У меня нет...	Я возьму / куплю....
большая пицца		
чёрный чай		
красная икра		
слнвочное масло		
свежая зелень		

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

III. _____/5 pts. Покупаем икру. *You are shopping online for the best deal on caviar. Review the list and prices and answer questions:*

№1  №2  №3 

Фото	Наименование	Вес Нетто	Цена
№1 	Икра лососевая зернистая (стекло)	350 г	500 руб.
№2 	Икра лососевая зернистая (жесть)	150 г	250 руб.
№3 	Икра лососевая зернистая Сахалин (жесть)	150 г	275 руб.

1. Какая банка икры самая дорогая?
   _____
2. Какая банка икры больше?
   _____
3. Какую банку лучше всего купить?
   _____

IV. _____25 pts. Сочинение. Essay. *You are about to call your Russian host family and discuss with them your food / drink preferences for specific meals of the day. What do you like to eat and drink*

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

in general and what food/drinks you do not like at all? Please write down a script for your call (greet a person, introduce yourself, and discuss your preferences).

OR:

IV. _____/25 pts. Диалог. *You can write a dialog instead. Your Russian roommate wants to cook a meal for you. You offer to go to the store to buy some necessary ingredients. Discuss with your friend what you need to buy and how much, suggest something, agree/disagree with your friend. The more you write while staying grammatically correct, the higher your grade will be.*

_____

_____

_____

_____

_____

_____

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 4-28: ***Scoring rubric for the oral exam (Section I)***

Студент (ка) _____					Chapter 11 Oral Exam
Fluency Vocabulary usage 25%	Content/Questions covered 25%	Language control Structure 20%	Pronunciation Comprehensibility 20%	Delivery /Confidence Ability to respond 10%	**Test points added:**
◯ **10** Excellent ◯ **9** Very good ◯ **8** Good ◯ **7** Satisfactory ◯ **6** Improvement needed	◯ **10** Excellent ◯ **9** Very good ◯ **8** Good ◯ **7** Satisfactory ◯ **6** Improvement needed	◯ **10** Excellent ◯ **9** Very good ◯ **8** Good ◯ **7** Satisfactory ◯ **6** Improvement needed	◯ **10** Excellent ◯ **9** Very good ◯ **8** Good ◯ **7** Satisfactory ◯ **6** Improvement needed	◯ **10** Excellent ◯ **9** Very good ◯ **8** Good ◯ **7** Satisfactory ◯ **6** Improvement needed	_____/25
_____X 2.5 =	_____X 2.5 =	_____X 2 =	_____X 2 =	_____X 1 =	_____/ 100%
Comments:					

10 Excellent Rich and extensive vocabulary with generally accurate usage (appropriate for this level); thorough treatment of skit/topic; (almost) entirely and easily comprehensible; grammatical structures are almost always correct (no systematic errors); no distracting pronunciation errors; responds to questions/comments adequately and appropriately; delivers the message in a confident, poised, enthusiastic/creative fashion.
9 Very good Extensive vocabulary with generally accurate usage (appropriate for this level); good treatment of skit/topic; (almost) easily comprehensible; grammatical structures are almost always correct (no systematic errors); a few distracting pronunciation errors; responds to questions/comments adequately and appropriately; delivers the message in a confident and composed fashion.
8 Good Occasionally lacks basic vocabulary, but generally good usage (appropriate for this level); grammatical structures are almost correct, but there are few systematic errors; few distracting pronunciation errors; needs some rephrasing/prompting, but responds adequately; delivers the message in an assured fashion.
7 Satisfactory Often lacks needed words, often inaccurate usage (below level) even of basic vocabulary; limited treatment of skit/topic; grammatical structures are correct only about half the time; several distracting pronunciation errors; needs frequent rephrasing/prompting, speech frequently hesitant; mostly routine phrases; at times difficult to understand; incomplete sentences; delivers the message in an uncertain fashion.
6 Improvement needed Mostly difficult to understand, even with instructor's/partner's prompting; weak treatment of skit/topic (way below the level); sentences mainly incomplete; lacks basic vocabulary; severe distracting pronunciation errors, jerky and slow language; delivers the message in an uncertain fashion.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 4-29: *Scoring rubric for the essay exam (Section II)*

Студент (ка) _____					Chapter __ Essay/Dialog **Test points added:**
Fluency Vocabulary usage 25%	Content Quality 25%	Language control Structure 20%	Comprehensibility 20%	Delivery /Creativity 10%	
◯ 10 Excellent ◯ 9 Very good ◯ 8 Good ◯ 7 Satisfactory ◯ 6 Improvement needed	◯ 10 Excellent ◯ 9 Very good ◯ 8 Good ◯ 7 Satisfactory ◯ 6 Improvement needed	◯ 10 Excellent ◯ 9 Very good ◯ 8 Good ◯ 7 Satisfactory ◯ 6 Improvement needed	◯ 10 Excellent ◯ 9 Very good ◯ 8 Good ◯ 7 Satisfactory ◯ 6 Improvement needed	◯ 10 Excellent ◯ 9 Very good ◯ 8 Good ◯ 7 Satisfactory ◯ 6 Improvement needed	_____
____X 2.5 =	____X 2.5 =	____X 2 =	____X 2 =	____X 1 =	_____/ 100%
Need to pay attention to:					
10 Excellent Rich and extensive vocabulary with generally accurate usage (appropriate for this level); thorough treatment of skit/topic; (almost) entirely and easily comprehensible; grammatica structures are almost always correct (no systematic errors); no distracting written errors; writes in a confident, poised, enthusiastic/creative fashion.					
9 Very good Extensive vocabulary with generally accurate usage (appropriate for this level); good treatment of skit/topic; (almost) easily comprehensible; grammatical structures are almost always correct (no systematic errors); a few distracting written errors; writes in a confident and composed fashion.					
8 Good Occasionally lacks basic vocabulary, but generally good usage (appropriate for this level); grammatical structures are almost correct, but there are few systematic errors; needs some rephrasing/prompting, but writes in an assured fashion.					
7 Satisfactory Often lacks needed words, often inaccurate usage (below level) even of basic vocabulary; limited treatment of skit/topic; grammatical structures are correct only about half the time; several distracting written errors; needs frequent rephrasing/prompting; mostly routine phrases; at times difficult to understand; incomplete sentences; writes in an uncertain fashion.					
6 Improvement needed Mostly difficult to understand, weak treatment of skit/topic (way below the level); sentences mainly incomplete; lacks basic vocabulary; severe distracting errors, jerky and slow language; writes in an uncertain fashion.					

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

English Translation and Answer Key of the Test

RUSN 1502 Экзамен Урок 11 Spring 2018

Chapter 11 Test

Имя, фамилия _____/100
Name of the student

I. Oral presentation _____**/20 pts.** Communicative task: *You and a friend (or any family member) are discussing an upcoming party on the phone and deciding what you need to buy, how much you need, and who is buying what.*

II. Грамматика. Language control.

1._____**/15 pts.** **Что вы обычно едите и пьёте на завтрак, на обед и на ужин?** *What do you usually eat and drink for breakfast, lunch, and supper?*
List 3 food items and 1 drink item for each meal. If you never eat your breakfast, add food items to your other meals. Answer in full sentences. Use verbs **есть** *(to eat) и* **пить** *(to drink).*

На завтрак
For breakfast

(3) Обычно на завтрак я ем яйца, бэкон и хлеб.

(1) Я всегда пью на затрак кофе.

На обед
For lunch

(3) На обед я часто ем курицу, рис и салат.

(1) Я люблю пить колу на обед.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

На ужин
For dinner

(3) На ужин я ем пиццу, гамбургер или бутерброд.

(1) Я часто пью шоколадное молоко или чай на ужин.

2._____/10 pts. Поговорим о еде. *Let's talk about food. Answer following questions in full sentences:*

1.Чего вы не любите есть? *What you do not like to eat?* Я не люблю овощи. Я не ем капусту.
2 .Какие напитки вы обычно пьёте? *What beverages do you usually drink?* Я обычно пью молоко, кока-колу или лимонад. Иногда я пью минеральную воду.

3. Какие фрукты или овощи вы едите каждый день? *What fruit and vegetables do you eat every day?* Я ем апельсины, яблоки и морковь каждый день. Я очень люблю виноград.

4.Что вы ели вчера на обед? *What did you eat for lunch yesterday?* Вчера на обед я ел (-а) лапшу, сыр, ветчину и салат. Я пил (-а) яблочный сок.
5. Какую еду можно купить в кулинарии? *What kind of food one can buy in ready-made dishes at a store?* В кулинарии можно купить готовый салат, бутерброды и пирожки.

3. _____/ 5pts. *Match containers to the items. Use proper case for the quantity of containers.*

пакет	**бутылка**	**коробка**	**банка**	**пачка**
carton	*bottle*	*box*	*can*	*packet*

Вчера мы купили бутылку красного вина, пакет молока, коробку шоколадных конфет,

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Yesterday we bought a bottle of red wine, a carton of milk, a *box of chocolates*
банку черной икры и три маленьких пачки чёрного чая.
can of black caviar, and three small packets of black loose tea.

4. _____/ 20 pts. *Provide appropriate case forms for the following food items:*

Food / **Еда**	**У меня нет...** *I do not have...*	**Я возьму / куплю....** *I will take/ buy...*
большая пицца *big pizza*	большой пиццы	большую пиццу
чёрный чай *black loose tea*	чёрного чая	чёрный чай
красная икра *red caviar*	красной икры	красную икру
сливочное масло *butter*	сливочного масла	сливочное масло
свежая зелень *fresh greens*	свежей зелени	свежую зелень

III. _____/5 pts. Покупаем икру. *Shopping for caviar. You are shopping online for the best deal on caviar. Review the list and prices and answer questions:*

№1 

№2 

№3 

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Фото *Photo*	**Наименование** *Item*	Вес Нетто *Net weight*	Цена *Price*
№1 	Икра лососевая зернистая (стекло) *Salmon caviar big grain (glass jar)*	350 г *8.8 oz.*	500 руб. *500 rubles*
№2 	Икра лососевая зернистая (жесть) *Salmon caviar big grain (tin can)*	150 г *5.29 oz.*	250 руб. *250 rubles*
№3 	Икра лососевая зернистая Сахалин (жесть) *Salmon caviar big grain from Sakhalin (tin can)*	150 г *5.29 oz.*	275 руб. *275 rubles*

1. Какая банка икры самая дорогая? *Which can is the most expensive?*

Самая дорогая банка № 3

2. Какая банка икры больше? *Which can is bigger?*

Банка № 1 больше.

3. Какую банку лучше всего купить? *Which can offers the best deal?*

Лучше всего купить банку № 1

IV. _____25 pts. Сочинение. Essay. You are about to call your Russian host family and discuss with them your food/drink preferences for specific meals of the day. What do you like to eat and drink in general and what food/drinks you do not like at all? Please write down a script for your call (greet a person, introduce yourself, and discuss your preferences).
OR:

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

IV. _____/25 pts. Диалог. *Dialog*. You can write a dialog instead. Your Russian roommate wants to cook a meal for you. You offer to go to the store to buy some necessary ingredients. Discuss with your friend what you need to buy and how much, suggest something, agree/disagree with your friend. The more you write while staying grammatically correct, the higher your grade will be.

Contributor's Questionnaire Responses

Section One: Test Planning and Writing

Why did you write the test? What were the purposes of the test?
I believe testing has a vital role in second language acquisition (SLA). Good tests show the effects of linguistic experiences on a language learner, and help learners demonstrate utilized primary linguistic data and get more motivated in learning the language. Tests help to show the efficacy of output, and tests cognitively and actively engage students. Tests assist instructors in navigating their students from the testing point to next knowledge. Tests are also skill checkpoints on meeting the course objectives for both students and instructors.

Both tests and instructor feedback promote students' noticing the conditions that are required for output to be useful in a real-life, organic environment.

How did you decide how many subtests to write? Was there some correspondence between a subtest and a course objective, or was there some other reason to create the subtests you did?
The number of subtests I write usually corresponds with a course objective, major points of a learned material, and appropriate communicative usage of the material. For the written part of the test here, the number of subtests corresponded with course objectives, and checks on major grammar points of Chapter 11. The oral test also corresponded to specific course content studied in Chapter 11.

Did you write as many items or test tasks as you needed, or did you write more than you needed, then pare the number down?
Once I had identified the main focus points of the chapter, I designed the items to check each student's language skills and level of proficiency. Each item in the test needed to make sense to a student and serve a practical or communicative application. Each item was analyzed, edited, and adapted to what I thought was students' possible levels of proficiency. Rather than

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

reducing the number of items, I revised items based on how learners handled the course content in class meetings before the test.

For example, with Section II.1 on food and drinks, I started out with fewer items, but after reviewing the chapter before the test and seeing students' demonstration of successful verb usage in class, I decided to add more items to this section. For section II.3, I created the "Containers" item to test learners' application of the accusative case to the verbs "buy" and "take," but after students seemed to be struggling with the extensive shopping vocabulary in class, I simplified the section to their level of proficiency. Finally, Section IV was initially created as a creative writing essay, but after observing great group-work in class, the option of creating a dialogue was added to give students the opportunity to practice a chat format (question-answer format).

How did you decide how many items to write for each subtest?
The number of items created corresponded to the purpose of checking students' proficiency. I considered the amount of vocabulary students could operate, their level of language control, points suggested by the material, the means of communication students can use, and what fun, useful ways students could apply the knowledge within a time-restricted format. I also tried not to overwhelm students during the test, and avoided redundant items to save precious time for responding to the test.

How did you decide how many test items to write in total?
It could be just one item, if the focus is on checking comprehension or composition skills, or it could be a multiple-choice format with many items, if, for example, only grammar is tested. For this particular test, the created items were essential for checking proficiency on specific course objectives.

Did you have one version of your test, or did you create a second equivalent version?
I used a second, equivalent version for the students who were absent during the exam time, or who were allowed to take a test in advance of the original test.

Were you concerned at how long the test would take to administer?
Yes, it is always a concern since we have 50-minute class or lab periods. Students must complete the test during that time.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Were you concerned at how long the test would take to score?
While I was assigned a teaching assistant to work with, who was thus a second rater, I was still concerned with how long the test would take to score. Part of my job is to supervise the teaching assistant, who must learn how to score tests.

Were you concerned how you might use the test items themselves for learner feedback?
During the Oral Exam, rapid responses with permitted self-correction showed primarily acquired knowledge. In other words, students received feedback right away either from a partner during the performance, or from the instructor right after. Written tests are slower to produce feedback since such tests rely on learners' conscious knowledge of the target language. The feedback from the instructor then needs to be more precise or detailed.

Students were provided with individual comments, and I created a handout for test corrections based on common mistakes. A second, equivalent version of the test was provided to students who missed the original test date, but had an excuse.

Did you consider having your students take your test on a computer? Why or why not? If not, what were your concerns?
Russian students are encouraged to write in cursive, so it was not an option. Writing in cursive is an important element to learning written Russian. It is also a very common way to socialize in Russia by writing notes in cursive. Otherwise, the test could be given on a computer.

Did you plan to allow learners to re-take a test for improvement? The same test, or a different test?
Students received a self-correction handout if they wanted to self-correct and improve their test grade, but did not have an opportunity to re-take the test itself. While it would be interesting to give them an equivalent test and see how they perform the same tasks, the program was very intense, and there was no time, so this was not an option.

What sources did you draw from for your test items?
The majority of the ideas for test items came from everyday communicative teaching, usage of authentic online materials, textbook exercises that target effective language learning, and years of experience of teaching the language and predicting what would work and what would probably not. The following standard guidelines for language learners are used for each assessment:

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

ACTFL OPI (Oral Proficiency Interview) guidelines for Russian: https://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012/russian

STARTALK NCSSFL-ACTFL Global Can-Do Benchmarks and Statements
https://www.actfl.org/publications/guidelines-and-manuals/ncssfl-actfl-can-do-statements

ACTFL Performance Descriptors for Language Learners
https://www.actfl.org/publications/guidelines-and-manuals/actfl-performance-descriptors-language-learners

Which test item formats do you prefer to use? Consider: Short answer, fill-in-the-blank, matching, cloze, performance test, etc.?
I prefer all of the above when they fit the purpose of the item.

What types of learner knowledge do you believe you are capturing in your test?
Section I. (Oral Exam) is targeting listening comprehension and speaking ability. Section II.1 (Vocabulary) captures the ability of learners to retain acquired basic vocabulary items. Section II.3 (Containers) captures the ability to interpret and identify a proper form. I really wanted students to try what native speakers do almost every day while shopping for food. This item was inspired by a funny situation in the Russian store with my own youngest daughter, who came from USA to visit with her grandma and had to shop for specific items and quantities/containers using proper Russian cases. I thought this application would be very useful if students were going to study abroad.

Section II.4 (Accusative and genitive cases usage) checks on appropriateness of the language use when writing. Section III (Shopping for caviar) checks on students' ability to scan material, extract the needed information, and make a decision while reading and writing. Section IV (essay or dialogue) provides students with a creative writing opportunity and allows me to check their level of language control and ability to apply knowledge.

What learner skills do you believe you are capturing in your test?
For the oral exam (Section I), students were given a communicative task and the opportunity to demonstrate their acquired level of language proficiency in potential real-life situations. For the remaining written parts of the test,

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

they were encouraged to demonstrate their individual skills of interpretation, language comprehension, andreading/extracting the necessary knowledge for analysis of given tasks. They were able to decide on the appropriateness and control of language needed, and were tested on their ability to put it all into writing.

If you wrote a performance test (where learners have to converse or present a topic, or write something above the sentence level), how did you get the ideas for what task learners had to do for the test?
For the performance part of the test (the oral exam, Section I), students were put into a situation that often happens in real life while studying abroad. They needed to discuss their food/drink preferences with their host families for specific meals of the day, to avoid issues like food allergies, strange looking or tasting food, or even to prevent a host family from overspending, since students are culturally aware that Russians are well-known for their hospitality and the way they accommodate their guests to the best of their ability. This communicative task was inspired by comments from our students who went abroad and lived with a host family.

For the written dialogue (Section IV), the task was adjusted to a real-life situation in which roommates want to cook their traditional or authentic meals for each other to share. Our students who study abroad often live in dorms, and often have international events and interactions with instructors and peers. Authentic food prepared for these events is unique to the culture in which students were raised. In class group-work, students of Russian try out ways to discuss what they will cook, what ingredients and food to buy, how much of that they will need, and restricting food items they probably won't find while abroad. It was a fun way to experiment with a real-life language application. Interestingly, almost half of the students went with this option on a test, rather than write an essay.

How did you get ideas on how to score learners' performances (the scoring criteria)?
Students were asked to write a sample that corresponds with a communicative task within the space provided, while staying grammatically correct. The scoring criteria for Section IV are more or less common for language essay grading. Students were initially given 25 points, with deductions of either a half or a whole point for each grammatical error.

For any errors related to material that had not been sufficiently covered in the period leading up to the exam, no penalty was given, though correct usage of said material potentially equated to an extra half-point, at the rater's discretion. Students were not penalized for using additional material they

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

had learned themselves, to encourage an interest in language application. However, a few learners made gross mistakes in a dish description, such as wanting to make a cheesecake, but then giving a dish name and ingredients in a way that native speakers would never understand it. In this case, no points were deducted, but concerns were communicated to the student.

Since it is a Beginning Russian course with the main focus on a communicative approach in learning and teaching, no essay rubric was created for the scoring of the essays or dialogues for this Chapter Test. I just wanted students to practice their writing. However, students were asked to provide enough writing to fill in the test essay format based on a communicative task provided.

If you used points on a scale for scoring learners' performances, how did you decide on how many points on a scale to use?

For the oral exam (Section I), I used a rubric (see Table 4-28) and points assigned were based on this rubric. I wanted to create an oral exam that was more student-friendly and less judgmental to emphasize positive and encouraging feedback. It has five points on the scale: excellent, very good, good, satisfactory, and improvement needed. Learning a language is a challenging task, and improvement is the goal for both the students and the teacher. Students responded to the rubric very well, and we discussed possible ways to improve their language abilities no matter what grade they received.

For a total 25 essay or a dialogue points (Section IV), the following criteria was taken into the consideration: 1.) Possible volume, 2.) Meaningful content, and 3.) Proper language command. For example, if the essay was too short, additional points were taken off. If students were able to write much more then asked while staying grammatically correct, and the content was good, we added some points (up to two possible points) to reward the student. For the language command and written grammatical errors, possible deductions were as follows: each roughly written mistake was penalized by a whole point, less-rough mistakes and typos cost the student a half-point. If the student made the same mistake up to three times, there was no further penalty. However, the fourth time the same mistake was made, it garnered an additional half or whole point deduction. After the grading, we provided the students with written comments and feedback for their essays and dialogues. See Table 4-29.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Were you concerned about whether you could get a colleague to help you score learners' task performances?
Since we had two raters for the test, one of whom was a teaching assistant I supervised, the discussion on grading took place prior the test, and we checked each other's consistency while grading.

Did you plan to give the scoring criteria to the learners for future self- or peer-assessment?
Yes, students were provided with Oral Rubrics scoring criteria and a Test Preparation Handout to study prior to the test.

Did you compare your test to the lessons that learners had?
Yes, the test format and style corresponded with lessons the students received, so no new shocking tasks or confusing extravagant items were created or used for the test.

Did you compare your test to the textbook or other materials learners used?
Yes. The test was compared to online materials and textbook/workbook exercises that I used during the learning of the chapter, and the exam itself used a format relatively familiar to the students.

Section Two: Test Administering and Scoring

Were you concerned about test security? What did you do to ensure test security?
Not really. The test was created in a way that made it hard to cheat, even with items like Section II.4 where students needed to change initial forms to the accusative or genitive case. Students produced either what they knew or did not know considering test time constraints or complexity of the items requiring them to know the material.

How did you deal with learners who missed the test, or who were late for the test?
If students had a valid excuse for their absence on a test, they were provided with an equivalent, second version of the test, which they had to complete at the Lab under proctored supervision. The time for completion was equal to the one in class.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

How did you prepare learners to take the test? Were there any test items or test procedures that were new to them?
I often create chapter review sessions prior to assessments. Students work in the Language Lab with authentic videos online, comprised of recorded native speaker audio, interactive grammar exercises, and overall practice for comprehension. They get an idea of how well they might do on a test by following a Lab manual and completing tasks during Lab time. Prior to Chapter 11 Tests (Oral and Written) we had a Lab scheduled, so students were able to practice and verify their skills. In addition to this, I usually set aside one class for a test review, where students have the chance to familiarize themselves with the test activities format, work on the hardest grammar points, and ask questions to clarify their concerns. I created a handout for such sessions prior to the Chapter 11 Test as well, so students could practice difficult language points based on their homework's common errors. We also worked on the vocabulary and structure for the essay portion. They were encouraged to ask questions, help each other, and lower their possible test anxiety level.

Did you write any of the test in the learners' first language? Why?
For students of Novice and Intermediate levels, I usually add a translation of instructions in English to avoid confusion or high level of anxiety.

For classroom tests: How did you accomplish scoring learners' tests?
I had an applied linguistics graduate student who was my teaching assistant at that time.

Did you write a test key?
There was no need. Although for the purposes of presenting this test in this book, an answer key has been provided above.

Did you go back and change your marks on previously scored tests in response to problems you found while scoring tests later in the process?
The two raters discussed some discrepancies and changed some marks on previously scored items in order to be consistent.

For performance tests: How did you accomplish scoring learners' performances?
Consider: Did you record learners' spoken performances to score later, or perhaps score a second time? Was recording equipment available?
Not for this test, since we took it in class with two raters available and feedback provided right after the performance. When we do recordings via

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

vocaroo.com or via a voice recorder, students would have the opportunity to identify their errors and recognize what they need to work on.

Did you score learners' spoken performances at the same time learners gave their performances? If so, did you have enough time to score?
Yes, we did score learners' spoken performances at the same time they were given. We had enough time to score since we had two raters.

Did you use any statistical reliability estimates, such as Pearson Product Moment to compare raters' scores on a performance test?
Several reliability analyses were provided by Matthew Hom, my teaching assistant, who statistically analyzed this test and reported the results as a project in one of his graduate classes.
He calculated reliability for the objectively scored written portion of the test using Cronbach's Alpha, with the result being a coefficient of .94, with a Standard Error of Measurement (SEM) of 2.23. for a test with 100 total possible points.

For the descriptive statistics on the objectively scored portions, he found that the mean was 87.5 and the median was 89.5, with a skewness of -.77, and a kurtosis of -.53. The negative skewness suggested a distribution toward the right, which is good for an achievement test done after instruction. Though the kurtosis, being negative, indicated a slightly "flat" distribution, the fact that the median was greater than the mean, and that the skewness was negative (which further means that most of the scores exceeded the mean,) means that most of the students did well on the test, which is ideal for a post-test, taken by learners after instruction. I really appreciate Mr. Hom's contribution in this regard.

Reporting scores

Did you report the scores to learners? If so, how did you report test scores to learners?
Yes, of course. Tests with grades and comments on each item (if necessary) were handed back to students.

What was your goal in reporting the test scores to learners?
My goal was to raise awareness and encourage self-investigation of errors, promote self-correction, and foster self-directed improvement.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you spend time explaining scores, or answering learners' questions about scores in or out of class?
Yes, we had a test results review session after students given back their tests.

How quickly did you report scores to learners? Was speed a priority?
As soon as we could. With two raters, it took us three days to grade both sections.

Section Three: Using Test Scores

Cut scores

How did you decide which scores were passing or failing scores (cut scores)? How did you decide which scores meant a specific grade on a test?
For this test it was a cumulative score that corresponded with the letter-score system as identified in the course Syllabus:

A 90-100
B 80-<90
C 70-<80
D 60-<70
F <60

The F score was a failing score.

For the Chapter 11 Test, different parts of the test were added up to create 100 points. The Oral Exam was assigned 25 points, the objectively scored Written Test made up 50 points, and the Essay/Dialog section was given the remaining 25 points.

How did you use learners' scores from this test?
Chapter 11 Test scores were used to calculate an overall Unit Grade Score for the Chapter learned. Homework, quizzes, Lab, Chapter test score, and extra-credit score (if any) were used to assign the Unit Grade for this Chapter. Chapter 11 Unit grades were posted on Blackboard for students' records.

What was the role of the test score in determining learners' grades? Consider: How much weight did you give your test? How did you decide?
The score on the test presented here was the part of the Unit Grade each student received. By the end of semester, students had six Unit Grades and were able to monitor their semester progress during the study.

We had six Chapter Unit Grades which were weighted as 35% of the total Final Grade for the semester. This test score was a part of the Chapter

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

11 Unit Grade, and was weighted as 30%, along with the other assigned work for the Unit Grade:

10 Homework assignments	40 %
4 Quizzes	24 %
1 Lab session	6 %
Chapter 11 Test	30%

Homework is essential for course practice and meeting course objectives, so any missing homework out of ten for the chapter cost students 4%. Students had four quizzes and a Lab session for 6% each. Oral and Written Tests were counted together for total of 30% of the Chapter 11 Unit Grade. 30% is a reasonable amount for the test that reflects learned material and helps to identify acquired language proficiency, but not too overpowering, if students worked well throughout the chapter. One of my students joked that "the Russian test is not the end of the world, but your life depends on it."

Were other measures used to decide learners' grades, besides your test?
Students' homework scores showed consistency of practice while learning Chapter material. I do not assign homework grades for beginners, but diligently provide them with feedback. Quiz scores showed how well students learned essential ongoing vocabulary. Lab scores showed students' ability to apply knowledge and to self-correct when necessary.

What was the relationship of the other measures to your test?
All other measures served to help students gain skills needed to perform well on the test, and had students work on resolving essential issues for their language proficiency tasks done during class meetings.

Did your test capture some knowledge, skill, or ability the other measures did not capture?
The test showed how well students might perform under pressure, a deadline situation, or under a higher level of anxiety. Some test items were easier for all students, and some were much more difficult and challenging. But even a test should extract a positive feeling of accomplishment and trusting oneself, so we cannot use only the most challenging items. Each item showed that if students were ready to go further, or wanted to try out something new, they have learned, both in and out of the classroom, a thought that might better express the meaning.

The test also showed if students were confused and needed more clarification on subject matter, or if they were overthinking or totally went blank due to frustration. The test really showed students' ability to

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

communicate and create with the language, and restructure and recover when they needed to deliver a message. Nobody gave up or fainted.

Reporting scores

How did you report scores to learners? Was timeliness of concern to you?
I usually grade everything as soon as I can, so students have almost immediate feedback as we go. Scores were reported to each student personally within a few days after completion, with comments either on specific items or for the whole test.

Did you hand the test back to learners? Did the learners get to keep the tests? Or did you take the tests back?
Tests were handed back, and students kept the tests as a review at the end of the semester and study material for the Final Exam preparation. If they decided to do any self-correction, they had to submit it with their original graded test, so the instructor can identify the areas that needed to be worked on, and after that, all of the students' work was returned.

Did you offer feedback to individual learners in addition to their test scores? Written? Orally? In or out of class?
I schedule a test correction overview during class time and offer feedback for all students, but can go into more detail or explanation per individual request during office hours, face to face, or in written form as a suggestion for next time language use.

Did you teach learners to interpret their scores?
We had a talk before the first test was conducted, explained what each symbol for catching a mistake is, so students were able to interpret their scores.

For performance tests, did you use the test criteria to help learners interpret their scores?
We used a speaking rubric for students to go by.

Using test scores

Do you feel you learned what you needed to learn from the test results about what learners could and could not do, or what they knew or did not know?
The test showed the benefit of training or practice for some students on properly identifying Russian cases, but also showed the overall

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

improvement on speaking proficiency and language comprehension for all students who took the test.

Did your test change how learners studied?
Yes. Students who struggled on previous tests were doing better on specific test items they found interesting or useful for a real-life language exchange. Some students told me that if they were to have a similar time with future "hard" items, for example on the final test, they know how to better deal with it.

Consider: Did you mention trouble points as general comments?
I did go over some items that caused trouble for students. Students were able to self-correct and provide answers. Group work helped those still confused or kept in the dark. This test also revealed something that many second language learners, including myself, are familiar with. I pointed out how overthinking or native language thinking might interrupt the proper usage of the second language they can and are able to produce otherwise. I asked them to keep that in mind if they wanted to work on test corrections.

Did learners ask you about the test itself (not the test scores) outside of class? If so, what did they want to talk to you about?
Students had a test review and practice handout with a description of the tasks they might have on a test, examples of pictures, and extra-credit possible tasks. If they had questions regarding the format, we discussed it in class.

Did learners' test scores change your teaching?
Test scores are indicators of what is working and what is not. Exams test both students and teachers. Overall grades for the Chapter test were reliable, but some test items were modified for future usage or even discarded due to students getting lower scores on them. A new teaching approach was applied to more complex test material to make it more accessible and clearer for students before the final test at the end of semester.

If you could turn back time, what would you change about your test? What would you change about your test administration?
First, I would seek a review on my test from other raters who teach and create tests for their own teaching. Second, I would spend more time on creating the test and formatting some items differently, trying to make them more engaging and useful for my students. Finally, I would reorganize my teaching time or spend more time on teaching specific material to ensure

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

students understood. I will definitely keep some items for future tests. Since we do not have a Russian store to go to here, it would be ideal to create an online test game that would allow students to virtually "shop" and "talk" to native speakers--customers and sellers--in order to complete tasks.

Section Four: Evaluating and Reviewing your Answers

To what extent do you think you've described recurrent patterns in your work with tests?

The main goal of the evolution and revised iterations of my exams is to test the success of methods I use to teach in the classroom, to help students learn organically and easily, and fully grasp a concept, both on paper and in speech. I base my exams in both individual and general acquisition practices. If something doesn't work, I change it. If it works, I build on it further.

The underlying focus is always on real-world application, emphasizing useful vocabulary and grammatical understanding. The ideal is to teach skills that the student can take with him/her to a Russian-speaking country and successfully apply the tested topics to everyday interactions with locals/peers. Allowing students to self-correct and work in groups to hone their studying skills covers bases that a typical standardized exam cannot, and marrying the concept of technical grammar to the practicality of spoken vocabulary ignites passion in all different types of learners, whether linguistically strong or not, to further pursue the Russian language and cultural knowledge acquisition.

To what extent is your test here an innovation, or something new, for you?

Since each student in a language class is a truly unique learner, each test I create and administer to our groups is exclusive, too. It is challenging to create one test that fits all, and with this Chapter test, students and instructors were able to see the progress made right away.

This test was a part of the larger course construction, so students were able to observe how well this piece of the big puzzle fits in for them. Every time, successful or not, the language acquisition is an innovation, the "aha!" moment for our students.

We need to design tests that are not destroying their amazing feeling of being curious and inspired by what they have learned. Through the tests, students also learn that a calm sea never made a skillful sailor, so it was working data for them to collect in order to evaluate the level of proficiency and go forward towards the improvement.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

La Historia De La Pola: An Achievement Test for Original Content-based Materials for Beginning Learners of Spanish

Yesica Amaya
Texas Tech University

Introduction

An Applied Linguistics M.A. student at a large public university in the southwest U.S., I am also a teaching assistant in our Spanish program. I have taught first, second, and third-semester Spanish courses. I am a native of Toca, Boyacá, Colombia. My research interests include the implementation of authentic materials in Spanish- and English-language beginner classes, second language testing, and flipped classroom teaching methods.

I created this test to check learners' achievement on a lesson based on authentic materials that I developed and adapted, designed for beginning learners of Spanish. The objectives of the lesson were to lead learners to develop different reading skills to understand general and specific information, to look for specific information while listening, to analyze patterns in written and oral texts to compare in which situations different time tenses were used, and to write a letter to a friend about Policarpa Salavarrieta's life. Policarpa was a Colombian heroine who fought and died for the independence of my country. I decided to use these materials because one part of the course objectives stated that students should learn about different cultures where Spanish is spoken. However, the time spent in the classroom to work on culture was very limited and usually culture was isolated from the language targets that students were learning. Thus, these materials not only offered students the opportunity to work on culture, but at the same time it allowed them to work on different language skills, to learn vocabulary, and to practice the specific language forms that they had learned during the course.

One of my general resources to develop this test was Gorsuch and Griffee's (2018) book, *Second Language Testing for Student Evaluation*

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

and Classroom Research. Based on the goals of the lesson, the constructs of this paper and pencil test were bottom-up skills for reading and listening, accurate use of the past and the present tense in continuous discourse, and textual competence providing details of historical events. To measure the constructs, the test was divided into two major segments. The first had three subtests ("parts") used multiple choice, matching, and cloze procedure items. The second major segment was a performance test which had a task and scoring criteria that I devised. I decided on a cut score after piloting the test. For the first part, the cut score was 14 (out of 22 possible points), and for the second part (the performance test) the cut score was 9 (out of 15 possible points).

It is important to point out that this test was low stakes since the course met for 14 weeks, with multiple lessons per week. This one test by itself would not determine if the learners passed or failed the class. The class was a hybrid course and used the flipped method. Therefore, students met three days per week in the classroom and there were no grammar explanations in class. It was students' responsibility to watch the tutorials through the software "My Spanish Lab" (https://www.pearsonmylabandmastering.com/northamerica/mylanguagelabs/) (Pearson Education, 2019) and, in general, the purpose of the class was to focus on communication to make students use what they had learned. The test takers were Spanish students of the lower level at my school (1507-2301-2302). The learners came from different undergraduate majors. Some of them were taking the class because they want to learn Spanish, and others because it was a requirement for their undergraduate degree. One striking this is that students have had widely different learning experiences since, in some cases, their families or close friends speak Spanish, and, in other cases, their only contact with the target language has been through formal lessons at school.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

The Test

Test: "La historia de la Pola"

Name: _____

Date: _____

Part 1

1. Listen to ten different sentences. For each sentence, circle the word you hear. You will listen to each sentence once:

a)	Nació	Nadó	Narró	Nombró
b)	Novena	Noventa	Nueve	Diecinueve
c)	Conoció	Comenzó	Contactó	Comprobó
d)	Cocinera	Costurera	Camarera	Comunera
e)	Ahorrar	Adaptar	Aplazar	Anticipar
f)	Veintiséis	Veintidós	Veinte	Veintitrés
g)	Lupa	Trucha	Lucha	Ducha
h)	Mayor	Mejor	Menor	Montón
i)	Valiosa	Valiente	Valerosa	Viviente
j)	Padre	País	Poetría	Patria

2. Match the period of Pola's life with the correct event by putting the letter in the chart. You may match only one event with each period of her life.

a) Infancia

___La Pola viajó a Santa Fé para aprender a trabajar como costurera.
___La Pola aprendió a leer y a escribir cuando era muy pequeña.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

b) Adolescencia

___La Pola escribió libros de historia acerca de la revolución.

___La Pola es recordada como una mujer valiente que murió por su país.

c) Adultez

___La Pola pensó en casarse con su amado Alejo Savarín.

___La Pola fue fusilada por la independencia de su patria.

d) Después de su muerte

___La Pola fue fusilada en Santa Fé por no trabajar por la independencia de Colombia.

3. Choose from the box the word that fills in the blank correctly. Each word may be used only once.

batalla / comenzó / costurera / admirada / recuerdan / nació / valiente / atacó / fusilada / independencia / trabajó / creció

La Academia Colombiana De Historia cuenta que Policarpa Salavarrieta (la Pola) (a)_____ en Guaduas el 26 de enero de 1795. En su infancia, la Pola aprendió a leer y a escribir y (b)_____ a interesarse en las actividades independentistas del país. En 1809 empezó a trabajar como (c) _____. Ese mismo año, Policarpa conoció a Alejo Sabaraín y empezó a ahorrar dinero de lo que ganaba en su trabajo para contraer matrimonio. Sin embargo, el matrimonio fue aplazado porque la Pola y Alejo decidieron unirse para luchar por la (d)_____ de Colombia. Un domingo del mes de julio se oyó del grito de independencia en Guaduas. Policarpa viajó a Santa Fé y (e)_____ en la casa de doña María Matea Martínez de Zaldúa. En Santa Fé, la Pola tuvo la oportunidad de ayudar con información

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

y contactos para la planeada revolución santafereña. Policarpa se destacó por sus labores de inteligencia durante la revolución. En 1817, Policarpa fue capturada y (f) _____ junto con su amado Alejo. Por ello, los colombianos la (g)_____ como una mujer joven y (h)_____ que murió por su patria.

Part 2 (Performance Test)

"La historia de la Pola" Performance Test
Name: _____
Date: _____

Imagine that one of your friends is going to Guaduas and s/he decides to stop by Policarpa's house, but he/she doesn't know who Policarpa was and why she is important in Colombian history. Write a letter telling him/her in your own words all the information you know about Policarpa. You can refer to your notes to answer this item.

Lubbock, 28 de Marzo de 2018
Querido (a) amigo(a) _____:
Escuché que estas en Guaduas y vas a visitar la casa de "La Pola". Quiero contarte su historia en esta carta.

_____
_____
_____
_____
_____
_____
_____
_____
_____

Te deseo unas felices vacaciones.

Sinceramente,

_____

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 4-30: *Scoring criteria for the performance test (Part 2)*

Points	1	2	3	4	5
Organi-zación	No writing or impossible to understand.	The organization of the events does not follow a logical sequence or are not enough.	The organization of the events are inaccurate organized most of the time.	Generally organized with only a few mistakes in the organization of the events.	Very organized. It explains the event in the correct order with a logical sequence.
Content	No writing, impossible to understand, does not include relevant information, or the student copied from other items.	The text only addresses few events about Pola's life with their own words and most of the information is inaccurate.	The text explains some of the events about Pola's life with their own words and there are only a few inaccurate information.	The text includes most of the events about Pola's life with their own words only a few inaccurate information.	The text address all the events about Pola's life accurately and with their own words.
Grammar	No control of grammar structures.	Sentences are incomplete and/or inconsistent	Fair control despite some confusion.	Effective use of grammar but some mistakes.	Expressions are clear and effective with rare mistakes.

Note. This represents the original scoring criteria. They were written in English to begin with.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

English Translation and Answer Key of the Test

Test: "La Historia De La Pola"
Name: _____
Date: _____

Part 1

1. Listen to ten different sentences. For each sentence, circle the word you hear. You will listen to each sentence once: Correct answers have been underlined.

a)	Nació	Nadó	Narró	Nombró
b)	Novena	Noventa	Nueve	Diecinueve
c)	Conoció	Comenzó	Contactó	Comprobó
d)	Cocinera	Costurera	Camarera	Comunera
e)	Ahorrar	Adaptar	Aplazar	Anticipar
f)	Veintiséis	Veintidós	Veinte	Veintitrés
g)	Lupa	Trucha	Lucha	Ducha
h)	Mayor	Mejor	Menor	Montón
i)	Valiosa	Valiente	Valerosa	Viviente
j)	Padre	País	Poetría	Patria

[Students hear:

a) Policarpa Salavarrieta "La Pola" nació en Guaduas el 26 de enero de 1795. *(Policarpa Salavarrieta "La Pola" was born in Guaduas in January 26, 1795.)*

b) En Guaduas, "La Pola" aprendió a leer y a escribir a los nueve años. (*In Guaduas, La Pola learned to read and write when she was nine years old.*)

c) En 1809 conoció a los hermanos Leandro y Alejo Sabaraín. (*In 1809, she met the brothers Leandro and Alejo Sabarín.*)

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

d) La Pola comenzó a ahorrar algún dinero de lo que ganaba de trabajar como costurera. (*La Pola started to save some money of what she received in her job as a seamstress.*)

e) Los acontecimientos revolucionarios obligaron a "La Pola" y a Alejo Sabaraín a aplazar el matrimonio para unirse a la revolución. (*The revolutionary events forced "La Pola" and Alejo Sabaraín to postpone their wedding to join the revolution.*)

f) El 22_ de julio, en Guaduas se supo la noticia del grito de Independencia. (*On July 22, the news about the call for the independence was heard in Guaduas.*)

g) El 3 de septiembre, Alejo Sabaraín y sus compañeros de lucha fueron detenidos. (*On September the third, Alejo Sabaraín and his battle comrades were arrested.*)

h) "La Pola" fue encerrada en el Colegio Mayor del Rosario. (*La Pola was arrested in Mayor del Rosario school.*)

i) Por ello, los colombianos recuerdan a "La Pola" como una mujer valiente que luchó y murió por la independencia de su patria. (*For these reasons, Colombians remember La Pola as a brave woman who fought and died for the independence of her homeland.*)

j) Por ello, los colombianos recuerdan a "La Pola" como una mujer valiente que luchó y murió por la independencia de su patria. (*For these reasons, Colombians remember La Pola as a brave woman who fought and died for the independence of her homeland.*)]

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

2. Match the period of Pola's life with the correct event by putting the letter in the chart. You may match only one event with each period of her life.

a) Infancia (*childhood*)

b) Adolescencia (*adolescence*)

c) Adultez (*adulthood*)

d) Después de su muerte (*after her death*)

___La Pola viajó a Santa Fé para aprender a trabajar como costurera. (*La Pola traveled to Santa Fe to learn how to be a seamstress*)

__A_La Pola aprendió a leer y a escribir cuando era muy pequeña. (*La Pola learned to read and write when she was young.)*

___La Pola escribió libros de historia acerca de la revolución. (*La Pola wrote books about the history of the revolution.*)

_D__La Pola es recordada como una mujer valiente que murió por su país. (*La Pola is remembered as a brave woman who died for her county.*)

__B_La Pola pensó en casarse con su amado Alejo Savarín. (*La Pola planned to marry Alejo Savarain.*)

_C__La Pola fue fusilada por la independencia de su patria. (*La Pola was shot for the Independence of her country.*)

___La Pola fue fusilada en Santa Fé por no trabajar por la independencia de Colombia. (*La Pola was shot because she didn't work for the Independence of her country.*)

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

3. Choose from the box the word that fills in the blank correctly. Each word may be used only once.

batalla / comenzó / costurera / admirada / recuerdan / nació / valiente / atacó / fusilada / independencia / trabajó / creció

La Academia Colombiana De Historia cuenta que Policarpa Salavarrieta (la Pola) (a) *nació* en Guaduas el 26 de enero de 1795. En su infancia, la Pola aprendió a leer y a escribir y (b) *comenzó* a interesarse en las actividades independentistas del país. En 1809 empezó a trabajar como (c) *costurera*. Ese mismo año, Policarpa conoció a Alejo Sabaraín y empezó a ahorrar dinero de lo que ganaba en su trabajo para contraer matrimonio. Sin embargo, el matrimonio fue aplazado porque la Pola y Alejo decidieron unirse para luchar por la (d) *independencia* de Colombia. Un domingo del mes de julio se oyó del grito de independencia en Guaduas. Policarpa viajó a Santa Fé y (e)) *trabajó* en la casa de doña María Matea Martínez de Zaldúa. En Santa Fé, la Pola tuvo la oportunidad de ayudar con información y contactos para la planeada revolución santafereña. Policarpa se destacó por sus labores de inteligencia durante la revolución. En 1817, Policarpa fue capturada y (f)) *fusilada* junto con su amado Alejo. Por ello, los colombianos la (g) *recuerdan* como una mujer joven y (h) *valiente* que murió por su patria.

The Colombian Academy of History explains that Policarpa Salavarrieta (la Pola) was born in Guaduas on January 26, 1795. During her childhood, la Pola learned to read and write and started to be interested in the Independence activities of the country. In 1809, she started working as a seamstress. During this year, she met Alejo Savaraín and started to save money to marry him. However, the wedding was postponed because "la Pola" and Alejo decided to join the battle for the Colombian independence. One a Sunday of July, the news of the call for Independence was heard in Guaduas. Policarpa moved to Santa Fé and worked in Mrs. María Matea Martínez de Zaldúa's house. In Santa Fé, "la Pola" had the opportunity to help with information for the revolution. Policarpa stood out due to her intelligence work during the revolution. In 1817, Policarpa and her love Alejo were arrested and shot. For these reasons, Colombians remember "la Pola" as a young and brave woman who died for her homeland.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Contributor's Questionnaire Responses

Section One: Test Planning and Writing

Why did you write the test? What were the purposes of the test?
I wrote this test as part of a course project that I was a requirement for a course in Second Language Testing that I was taking during the first year of my master's studies. In a previous course, I had created a unit with authentic materials which motivated me to write this test for those specific materials. In that way, I was going to be able to have a test to measure if the students had achieved the objectives proposed in the unit after applying it.

How did you decide how many subtests to write? Was there some correspondence between a subtest and a course objective, or was there some other reason to create the subtests you did?
I decided the number of subtests based on the objectives that I had in the unit. Therefore, I had objectives such as developing bottom-up skills for reading and listening, accurately using the past and the present tense in continuous discourse, and developing textual competence providing details of historical events. Thus, I had one subtest for each of the objectives.

Did you write as many items or test tasks as you needed, or did you write more than you needed, then pare the number down? How did you decide which items or test tasks to keep? How did you decide which items or test tasks to discard?
In the beginning, I wrote more items than what I needed because I was not sure if they were going to work as expected. Then, I piloted the test and calculated the item facility (IF) and the difference index (DI). IF has to do with how easy the group found the item. DI has to do with finding items that the learners improved on overtime, or in my case, specific items that showed a difference between a master and non-master group.

At that time, I hadn't the opportunity to apply the materials to pilot the test. Therefore, I had to look for two groups that could function as masters and non-masters. I was expecting that the master group would have acquired in their Spanish classes the constructs that my test was measuring since they were at a more advanced level. The non-master group was formed by students who were complete beginners, so it was expected that they wouldn't have those constructs yet.

After the two groups took the test under the same conditions, the results were compared and the descriptive statistics were calculated. However, the information that was crucial to establish what items needed to be revised or

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

excluded from the test was the DIs and the Ifs for each item. The items with a low DI (items showing no apparent differences between groups) were revised or discarded to create a revised version of the test.

How did you decide how many items to write for each subtest?
Since my test was designed to measure if the learners would achieve the objectives proposed in the unit and the unit was not very long, I did not want my test to be too long either. However, I still needed a good number of items, especially for the piloting, since I didn't know if all of the items were going to work. My professor suggested that I include more than ten items for the subtests of the listening, the fill in the blank items, and the matching. Since the texts about Pola were short and covered only four specific periods of her lifetime, it was not possible to add more than four items for the matching subtest (subtest #2). For this reason, my professor suggested I have more options for the students to choose for each period of time.

How did you decide how many test items to write in total?
I decided based on the idea that I need to have at least eight working items in each subtest. Thus, I could write ten to twelve items, so if I had to remove some after calculating the IF and the DI, I could still have a number around eight.

Were you concerned at how long the test would take to administer?
Yes. I did not want my test to be very long because it was for the unit that I created and it could be considered a low-stakes test. Also, I was concerned about the time that students would have to spend since their participation was voluntary and they most likely didn't want to spend a long time answering a test that did not count for their grades. Furthermore, I needed to find a schedule that worked for all of them which made it harder because their schedules were different.

I did have the opportunity to implement the Pola unit again in one of my classes to develop a summative evaluation study for an M.A. course I was taking, so I used the test as a pre- and post-test again, with the same group of students. This time, I was also concerned about the time because I was given two days to apply the units, so I didn't have a long time to use for the test, but I know that I needed to give the students enough time to answer.

Were you concerned at how long the test would take to score?
Although this was not really something that I thought about, as teachers we are always concerned about how to manage our time and scoring is one of the things that takes the most time. I was in a normal semester as a graduate

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

student, taking classes and teaching, so I definitely did not want to spend hours and hours grading the test. However, my biggest concern was about the time that it would take to the student to answer, so I didn't pay much attention to how long the test would take to score.

Did you consider having your students take your test on a computer? Why or why not? If not, what were your concerns?
I didn't consider this. Since the beginning, I planned my test to be paper-based because it is easier to implement. If I had the students taking it on a computer, I would have to be worried about booking time in the language lab and making sure that everything was working appropriately. I would have been nervous about something going wrong, such as the computer not working well, so for comfort and because it is easier to implement, I chose paper-based.

Did you consider allowing learners to use additional sources such as dictionaries, or their notes, while taking the test?
I did not consider this option because I wanted to test what student had learned during the implementation of the materials and I was not sure in what cases it would be appropriate to give students access to additional sources during the test.

What sources did you draw from for your test items?
The test items ideas were based on what students did in class with the materials and the course objectives. I tried to make sure that the test had a positive washback and that it was measuring what I wanted to measure. Therefore, I first looked at the objectives of the unit and then to what students did in the units, so I could find items ideas. For example, there were some listening activities where students had to listen for specific words as they watched a video. Also, they listened to the story being narrated and then they had to write on the board the words they remembered. This game was the idea for the first subtest.

Which test item formats do you prefer to use? Consider: Short answer, fill-in-the-blank, matching, cloze, performance test, etc.?
I prefer to use performance tests, matching, fill in the blank, and multiple choice depending on what I want to measure.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

What types of learner knowledge do you believe you are capturing in your test? How does that change with test item types you used on the test?
The leamers' knowledge that the test was designed to capture was language knowledge, more specifically the use of present, preterit, and imperfect grammatical forms, as well as vocabulary. I think the test also captured textual and functional knowledge since students had to write a letter where they had to organize the information accurately to be able to tell Policarpa's story to their friend.

What learner skills do you believe you are capturing in your test? How does that change with test item types you used on the test?
The leamers' skills that the test was designed to capture were the ability to use button-up skills to listen and choose the correct word (multiple choice) and to read and match the appropriate event with the period of time (fill in the blank). Use accurately the present and past tense to narrate historical events in an email (performance test).

If you wrote a performance test (where learners have to converse or present a topic, or write something above the sentence level), how did you get the ideas for what task learners had to do for the test? Consider: From tasks learners do in class? From tasks they have do to in real life?
I have one performance test as a subtest (Part 2). The idea came from what students did in class with the materials since, based on what they learned, they had to write an email to a friend explaining about Policarpa's life.

How did you get ideas on how to score learners' performances (the scoring criteria)?
The ideas to score the performance test came from a course in second language testing that I was taking. In an M.A. testing class, there was a chapter where we read about how to write scales and we discussed which type would be better based on what we wanted to measure in the performance test. In my case, I chose an analytic scale because I wanted to clarify which areas I was going to focus on when scoring and I planned to have a rating scale for each area that allowed me to give a very specific score. See Table 4-3●.

If you used points on a scale for scoring learners' performances, how did you decide on how many points on a scale to use?
I used points based on examples of what we read in the M.A. testing class. Also, I remember we discussed that the greater the number of points on a scale, the harder it is to grade since it may become difficult to differentiate

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

one point from another (one higher performance than another). For this reason, I have three criteria with a rating scale from one to five for each. The description of each point in the rating scale needed to be very specific, so any teacher could understand it and when scoring it would be clear what each one meant. I also revised my scale during the piloting stage.

Were you concerned about whether you could get a colleague to help you score learners' task performances? Consider: Did you have ideas about how to ensure score consistency, such as having two scorers (you and a colleague), or scoring learners' performances by yourself on two different occasions?
I did know that it may be useful to have another college to help me score the learners' performances. However, since it was not a high-skates test, I decided to do it alone. In my M.A. testing course, I learned the technique of scoring the tests myself on two different occasions to check consistency in scoring.

Did you plan to give the scoring criteria to the learners for future self- or peer-assessment? Did you make another, perhaps simpler or shorter, version of the scoring criteria for learners to use?
I did not give the score criteria to the learners because my time to work on the materials and the test was very limited since it was not part of the regular curriculum using in the course.

Did you seek help from a peer to clarify what your test items or tasks were measuring?
Yes. I worked with my professor and one of my graduate student classmates to make sure that the items I had created were measuring what I wanted.

Did you ask another teacher to compare your test items with what you said you wanted to measure?
Yes, I asked one of my classmates from the testing course. She was also a language instructor in the department.

Did you make any changes to your test or items as a result of your colleague's feedback?
Yes. At the beginning when I was writing the test, I had thought of a different item format, but after discussing it with other people, I realized that it was better to use matching.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you compare your test to the lessons that learners had?
Yes. Before writing the test I looked at the unit I had created, so I could have a test that reflected what students did in class.

Section Two: Test Administering and Scoring

Were you concerned about test security? What did you do to ensure test security?
I was not really concerned about it. I just made sure that the students didn't have any notes or other materials when they were taking the test and I encouraged them to do the best that they could.

Were your tests photocopied, or did learners see your test items or test tasks another way, such as on the blackboard?
The tests were photocopied and each student received one copy.

How did you deal with learners who missed the test, or who were late for the test?
During the piloting, the students who showed up were the ones who volunteered, so I did not have any problems with time or absences. During the implementation of the unit and test the following semester, there were some students who did not attend the class on the day of the pre-test and the first lessons on the unit. Even though they then still took the post-test, those students' scores were not analyzed with the others since they had missed part of the lessons and did not have a pre-test I could use to compare with their post-test scores.

How did you prepare learners to take the test? Were there any test items or test procedures that were new to them?
For the piloting, when I had the master and non-master groups, most of the information of the test was new since it was about the authentic materials that neither group had studied. Therefore, I had to use one of the audio recordings of the authentic materials I made and play it to both groups twice and allow them to take notes, so they could be familiar with the topic and be able to answer the items.

I had supposed that the master group could answer the items if they had information about the text. For the non-master group, they had access to the information under the same conditions of the other group, but since they were lower level and without the skills I was trying to measure, even with the listening section that they had about the story, they could not answer all the items correctly.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Regarding the test items, none of the formats were unfamiliar. Usually, the exams that they take in the Spanish class include similar test item formats.

Did you pilot your test? Do a trial run? Did the pilot result in any changes to the final version of your test?
Yes. I piloted the test and based on the results, I created a revised version that I used when I implemented the materials. After the piloting, I realized that there were some minor changes that I could do to improve the test.

Did you write any of the test in the learners' first language? Why?
I wrote the instructions in English because I wanted students to be clear on what they needed to do to answer each subtest.

For classroom tests: How did you accomplish scoring learners' tests? Consider: Did you hide learners' names as you scored?
The three first subtests were objectively scored, meaning that there was only one possible answer, so I did not score them a second time.

Did you put learners' responses to items into a spreadsheet for further analysis? Did that process help you catch scoring accuracy problems? Or problems with bias?
I put the students' scores into a spreadsheet for further analysis. I did this process twice, one for the piloting of the test and the second time with the final version after the student studied the materials. During this process, I calculated the IF (Item Facility) of each test (pre- and post-) and then calculated the DI (Difference Index) to know if the items were working well and if the student had improved their scores. During this process, I didn't find any accuracy problems.

For performance tests: How did you accomplish scoring learners' performances? Consider: Did you record learners' spoken performances to score later, or perhaps score a second time? Was recording equipment available?
For the fourth subtest, the performance test, I did score the learners' answers twice. For that, I tried not to look at the students' names and I put the first score on the back of the sheet, so I could not see it when I was grading the second time. I did not find many differences between the scores. They usually were the same the second time around.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Do you think your test was reliable? What did you do to check?
I consider my test to be reliable because when I was designing it, I received suggestions about the items and my constructs from my professor and one of my classmates. After that, I made some changes to the test such as making sure the there was only one possible answer for the fill in the blank part, that the instructions were clear, and that the test was organized in a way that was easy for the student to respond.

During the piloting, all the students took the test under the same conditions and had enough time to answer the test. After that, I made revisions to the test based on the IF (Item Facility) and DI (Difference Index) that I calculated on the pilot test data. During the implementation stage (after the pilot), I was able to see how the test worked with the actual materials. I made further revisions after calculating the IF and DI again.

Likewise, during the scoring of the performance test, I was the only rater but I put the test aside and scored a second time.

Reporting scores

Did you report the scores to learners? If so, how did you report test scores to learners?
I did not have time to report the scores to the learners, unfortunately.

Section Three: Using Test Scores

Cut scores

How did you decide which scores were passing or failing scores (cut scores)? How did you decide which scores meant a specific grade on a test? Did you consult a testing book or think of previous coursework you had to determine cut scores?
For the cut scores, I used the contrastive group methods that I had learned in my testing class and from the testing book that we read. After scoring the pilot test, I put the results in a histogram and compared the master and the non-master group to know where they overlapped. In that way, I was able to see where the lines intersected and that was the cut score. When I organized the scores in the histogram, it showed that it overlapped at three different points. I decided to choose the higher (14 rounded up from 13.5 out of 22 possible points) because I considered that 14 could represent a "C-student" (minimally competent) since that was one of the highest scores that students in the non-master group had.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

For the performance test, the contrastive group method would not work, so I planned to use the direct consensus method, which consisted in showing the test task and the scoring criteria to another colleague to examine what level on the scoring criteria, in terms of the test task, students should have been able to reach in order to pass the test. I didn't specify grades such as A, B, F, etc., because the test was not part of the course that I was teaching, so I didn't have to report them in that way.

Using test scores

Do you feel you learned what you needed to learn from the test results about what learners could and could not do, or what they knew or did not know?
I think that the test gave me a good idea of what the learners were able to do based on the constructs that I stated ahead of time, while designing the test. It showed me that most of the students improved after studying the authentic materials, so I think the unit was effective to help learners to develop the specific skills I had in mind.

Did your test change how learners studied?
Consider: Did you use learners' scores to find out if your test caused washback?
Did you use particular item types or a performance test to change learners' practices or support their learning? Did their scores indicate they had changed their learning practices?
I am not sure about it. The unit I implemented had a completely different methodology than the one we usually use in the classroom. Therefore, the students had to read authentic materials, discuss content with their partners, and complete meaning-based matrices. All of this is something that they don't usually do. However, since it was a two-day lesson, I cannot say if it changed the students' practices. During that period of in-class work, learners did change how they studied, but I don't know to what extent it affected their study habits outside the class.

Did you spend time going over the test in class?
Before the test, I explained to the students what they were supposed to do in each subtest and I also allow them to ask questions if they had doubts as they were answering.

Did learners' test scores change your teaching?
It didn't because I didn't continue working on the unit. I had to move to the normal class schedule, and although I tried to include readings when I had

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

the chance because I saw that it helped students a lot, I had to follow a very specific schedule and method that didn't allow to implement those types of materials very often.

If you could turn back time, what would you change about your test? What would you change about your test administration?
I would give the learners feedback about their performance on the test and I would also be clearer about the purpose of the test since it was not part of what we usually do in class. I would probably show them the scale that I used to grade the performance test and the criteria that it had, so they could be more prepared and know what I was expecting to find.

Section Four: Evaluating and Reviewing your Answers

Please read through your answers to the items and answer the following:
To what extent do you think you've described recurrent patterns in your work with tests?
I think I have described recurrent patterns related to the design of the test and the analysis of the scores comparing the pre-test and the post-test. Also, the importance of considering the objectives and what the students did in the lessons as well as piloting and revising the test items.

To what extent is your test here an innovation, or something new, for you?
For me, it is an innovation, first of all, because it was the first time I had designed a test taking into consideration all the aspects that I mentioned through this questionnaire. In the past, I was not familiar with the procedures for designing a test and how important it was to make sure that it was measuring what I wanted to measure. Also, it is an innovation in some way for me because even though I designed a very small unit, I learned that it is still important to have a reliable test to measure what the students had learned with those materials.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

A Final Project Performance Test for a Spanish Conversation Class at a Korean University

Maria Teresa Martinez-Garcia
Hankuk University of Foreign Studies

Introduction

My name is Maria Teresa Martinez-Garcia and I am an assistant professor of Spanish at Hankuk University of Foreign Studies (Seoul, South Korea), where I teach Spanish classes. The classes range from communicative/elementary Spanish to advanced debate or essay writing. My background is in linguistics (Ph.D., University of Kansas), and my dissertation research (Martinez-Garcia, 2016) took a psycholinguistic approach to understanding bilingual activation, by exploring how differences in stress placement between English-Spanish identical cognates affect how adult learners of Spanish use stress as a cue for word recognition. My research interests include bilingualism, second language acquisition, speech perception and production, and pedagogical approaches to teaching pronunciation in the foreign language classroom.

The test I present here was used with my intermediate conversation class. Hankuk University of Foreign Studies (HUFS) is a private research university, which specializes in foreign language education, offering 53 different language courses. The Spanish department is one of the largest departments at the university (together with the English and Chinese departments), with about 500 students enrolled seeking to graduate with a major or a minor in Spanish. Based on department policy, the intermediate conversation class has been created to help students reach a B1 level in Spanish speaking, following the standards of the Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001). The class consists of 20 students who weekly meet for 2 hours, for a total of 16 weeks. When students take this class, they have normally taken two beginner conversation classes and most of the grammar and composition classes. The latter amounts to four semesters of grammar courses and two

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

semesters of composition. Apart from these classes, which are mostly taught in Spanish, these students have already taken most of the core courses in the Spanish department (e.g., History of Spanish Culture or Latin American Literature), which are normally taught in Korean, most learners' L1. In order to graduate with a major or a minor in Spanish, students need to complete six semesters of conversation in Spanish, which is the focus of the test presented here.

One of the challenges of teaching in South Korea is to get students to participate in conversation classes. My feeling is their grammar is good, and that they can write complex essays about almost any topic, even from their first semester of formally learning Spanish. However, given the East Asian culture in which they have grown up, they are mostly used to classes in which the teacher/professor lectures and they only need to take notes, so it is normally difficult to change the dynamics of the class so that they are the ones doing most of the talking. With this project, I wanted students to do some real talking (they had the opportunity to plan and prepare out of class), but still let them decide about what they want to talk, to increase the communicativeness of the project.

The Test

Learners receive the test, the teacher's grading rubric (see Table 4-31), and the student's grading rubric (see Table 4-32), in Spanish. The instructions, together with the rubrics, are handed to students in the middle of the semester, to make sure students enough time to complete the project. After discussing the details of the test in class, students form groups and start their independent work preparing the final project (meeting to discuss the topic to be covered, recording the video, and preparing their final presentation). During this time, students are encouraged to attend my office hours to get help on the different steps of the project. Office hour topics range from brainstorming possible ideas of topics to drafting their final presentation.

When scoring the final projects (video and oral presentation) it is really important for me to see whether students would meet the requirements to pass the DELE B1 oral exam (see Baztán, Torrecillas, Cuadrado, Guerrero, & Molero, 2015). For that reason, and knowing how familiar Korean students are with memorizing information, I include an extra question for which they could not prepare and that would give me more information regarding their real level of proficiency in the language in terms of understanding and speaking. Moreover, I evaluate their complete performance using the grading scales created by the Instituto Cervantes to evaluate the speaking section of the B1 DELE exams, and which focus on

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

coherence, fluency, linguistic scope (i.e., lexis), and correctness (see Instituto Cervantes, 2019).

PROYECTO FINAL

El Proyecto Final se debe completar en parejas o grupos de tres personas (máximo). Para completar el proyecto final, debéis responder a la siguiente pregunta:

¿Qué significa esta universidad para ti?

El proyecto final evaluará tanto la gramática/vocabulario como la presentación y la originalidad, y consiste de dos partes, relacionadas entre sí.

1. **PARTE 1 (vídeo: 40 puntos)**

   Instrucciones: Vais a responder a la pregunta principal: ¿Qué significa esta universidad para ti? grabando un vídeo en el que mostráis porqué la universidad es importante para vosotros.

   - Detalles del vídeo: El vídeo debe mostrar imágenes que representen vuestra respuesta a la pregunta. No debe contener voz (está bien incluir algunas palabras (en español) o sonidos de voces de fondo). Podéis incluir música si queréis (los ejemplos os darán una idea).
     - Durante la **presentación** en clase es cuando incluiréis la "voz" y las palabras para responder a la pregunta y explicar las imágenes.
   - Duración del vídeo:
     - Grupo de 2 personas: Vídeo de unos 4 minutos.
     - Grupo de 3 personas: Vídeo de unos 6 minutos.

   Ideas: A la hora de grabar el vídeo pensad en temas como: la enseñanza, el crecimiento personal, la integración, nuevas oportunidades, las relaciones personales, la amistad, el futuro (laboral y personal), etc.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Ejemplos de vídeos:

- ¿Qué significa la verdadera amistad para ti? https://www.youtube.com/watch?v=ozTrgZdntwI
- ¿Qué significa la educación para ti? https://www.youtube.com/watch?v=Un5msddQl6U

2. PARTE 2 (presentación: 60 puntos (50 presentación + 10 comentarios de compañeros))

Instrucciones: En grupos, vais a presentar el vídeo. No se puede leer (aunque podéis tener notas con vosotros para ayudaros) y todos los miembros del grupo deben hablar más o menos durante el mismo tiempo.

Formato:

1. **Presentación** (unos 2 minutos): Presentación de los integrantes del grupo y de la motivación del vídeo (sobre qué temas trata y por qué pensasteis que eran temas importantes).

2. **Vídeo** (duración del vídeo): En este punto, debéis poner la voz en *off* al vídeo. Como narradores de la historia, debéis explicar qué está pasando en el vídeo.

    a. No es necesario hablar durante los 4 (o 6) minutos del vídeo, pero sí la mayor parte del mismo.

3. **Conclusión** (unos 2 minutos): Resumen del proyecto, incluyendo cómo ha cambiado vuestra perspectiva de la universidad (o no) desde que sois estudiantes.

4. **Pregunta** (no preparada): Pregunta no preparada, pero relacionada con el proyecto presentado. Todos los miembros del grupo deben responderla individualmente. Ejemplos de preguntas:

    a. ¿Por qué decidiste estudiar en esta universidad?
    b. ¿Volverías a tomar la decisión de estudiar en esta universidad?

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

c. ¿Qué cambiarías del tiempo que has pasado en esta universidad?

Table 4-31: ***The teacher's grading rubric Spanish version***

RÚBRICA DE LA PRESENTACIÓN (profesora)

	Supera las expectativas / A	Cumple las expectativas / B	En progreso / C	No cumple las expectativas / D-F
Léxico	Utiliza de forma apropiada y con efectividad el vocabulario necesario.	Utiliza de forma apropiada y bastante efectiva el vocabulario necesario.	Utiliza el vocabulario apropiado la mayor parte del tiempo; palabras en otros idiomas.	No utiliza el vocabulario de forma adecuada; el uso de otros idiomas afecta la comunicación.
Gramática	La gramática básica está toda perfecta.	La gramática básica está casi perfecta.	La gramática básica presenta errores importantes; influencia de otras lenguas.	La gramática presenta muchos errores; influencia de otras lenguas.
Fluidez	No existen problemas de pronunciación importantes. No hay demasiadas pausas y el discurso el fluido y coherente.	Hay pocos problemas de pronunciación (4). Hay pausas y muestra dudas, por lo que el discurso no es fluido.	Hay bastantes problemas de pronunciación (+4). Hay pausas y muestra dudas, por lo que el discurso no es fluido.	Los problemas de pronunciación afectan la comprensión. Existen pausas largas, dudas, y el discurso no es fluido.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Comprensión y producción	Completa la tarea con éxito. La gramática no se limita a estructuras básicas. Hay comunicación e interacción entre ambas partes.	Buen entendimiento de la tarea y utilización del lenguaje apropiado. Hay comunicación e interacción entre ambas partes.	Entiende la temática del proyecto, aunque la tarea no está completa o no es lógica. La participación entre los miembros no está compensada.	El grupo no entiende el proyecto y la producción está incompleta o no tiene lógica. La información proporcionada es demasiado básica o incluso inexistente.
Originalidad	El proyecto es único, no se parece a los demás. Muestra creatividad, es único y fresco.	El proyecto funciona, pero no es único ni original. Tiene bastantes componentes repetitivos.	El proyecto no está bien organizado. Tiene muchos componentes repetitivos.	Los estudiantes apenas prepararon el proyecto final.

Most universities in South Korea follow a relative grading system, which is a government-imposed grading curve. While it depends on schools, this grading curve demands that at least 30 percent of students are destined to receive a grade of C+ or lower. So, even when students deserve to get a higher level, the grading system at the university does not allow professors to give more than a certain percentage of As, of Bs, and of Cs. This means that the difference between a student getting and A and a student getting a C may be a couple of points (depending on how good the class is, someone with 91/100 points at the end of the semester may end up getting a C).

Thus, in Korea, getting a C is almost as bad as failing the class. Students often request to know in advance how exactly each individual project is going to be evaluated. Having these rubrics at hand help them identify in which aspects they may be weaker and how they may be able to improve their performance on the project and, hopefully, their grade.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 4-32: ***The students' grading rubric Spanish version***

RÚBRICA PARA EVALUAR LAS PRESENTACIONES (estudiantes)

Tu nombre: _____Presentación número: _____

1. Del 1 al 10, siendo 10 la máxima nota, ¿cómo evaluarías la originalidad del vídeo de este grupo? 1 2 3 4 5 6 7 8 9 10 2. Del 1 al 10, siendo 10 la máxima nota, ¿cómo evaluarías la claridad de la presentación de este grupo? 1 2 3 4 5 6 7 8 9 10 3. Del 1 al 10, siendo 10 la máxima nota, ¿cómo evaluarías la preparación de este grupo? 1 2 3 4 5 6 7 8 9 10
¿Tienes algún comentario más para el grupo?

English Translation of the Test and Scoring Rubrics

FINAL PROJECT

The Final Project must be done in pairs or groups of three persons (maximum). To complete the project, you need to address the following question:

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

What does this university mean to you?

The Final Project will evaluate both the use of grammar/vocabulary, as well as your presentation skills and the originality of the project. It consists of two parts, related to each other.

3. **PART 1 (video: 40 points)**
   Instructions: You are going to answer to the original question: What does this university mean to you? recording a video in which you show why the university is important for you.
   - Details about the video: The video must show images that relate to your answer to the question. You cannot record yourself speaking in the video (although it's ok to include some words (in Spanish) or background voices). You can include music if you want to (the examples will give you an idea).
     - During the **presentation** in class, you will include the voice-over and the words to answer to the question and to explain the video.
   - Duration of the video:
     - Group of 2 persons: Video of about 4 minutes.
     - Group of 3 persons: Video of about 6 minutes.

   Ideas: When considering how to address the main question of the Project, think about the following topics: teaching, personal growth, integration, new opportunities, personal relationships, friendships, the future (work-related or personal), etc.

   Examples of videos:
   - What does real friendship mean to you? https://www.youtube.com/watch?v=ozTrgZdntwI
   - What does education mean to you? https://www.youtube.com/watch?v=Un5msdd[redacted]l6U

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

4. **PART 2 (in-class presentation: 60 points (50 presentation + 10 other students' feedback))**

Instructions: In groups, you are going to present your video. You cannot read (although you can have notes with you as an aid) and all the members of the groups need to participate equally in the presentation.

Format:

5. **Introduction** (about 2 minutes): Introduction of the members of the group and the motivation of the video (talking about which topics it covers and why you thought they were important.

6. **Video** (duration of the video): At this point, you need to include the voice-over to the video. As narrators of the story, you need to explain what is going on in the video.
    a. You don't need to speak during the 4 (or 6) minutes of the video, but during most of that time.

7. **Conclusion** (about 2 minutes): Summary of the Project, including how your perspective of the university has change (or not) since you are students.

8. **Question** (not prepared): Question not prepared but related to the Project presented. Everybody in the group must answer it individually. Examples of questions:
    a. Why did you decide to study in this university?
    b. Would you make the same decision to study at this university?
    c. What would you change about the time you have spent in this university?

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Below please find the English translations of the test sheet, the teacher's grading rubric (Table 4-33) and the students' grading rubric (Table 4-34).

Table 4-33: ***The teacher's grading rubric English version***

RUBRIC TO EVALUATE PRESENTATIONS (teacher)

	Above expectations / A	Meet the expectations / B	In progress / C	Does not meet the expectations / D-F
Vocab-ulary	Employs appropriately and effectively the required vocabulary.	Employs appropriately and quite effectively the required vocabulary.	Employs the appropriate vocabulary most of the time: words in other languages.	Doesn't use the vocabulary in a proper way; overuse of foreign languages, which affects communication.
Grammar	Basic grammar is perfectly used.	Basic grammar is almost perfectly used.	Basic grammar present important errors; influence from other languages.	Grammar presents a lot of mistakes; influence from other languages.
Fluidity	There are no important problems in pronunciatio n. Barely any pause and fluid, coherent discourse.	There are some errors in pronunciation (4). There are some pauses, and shows doubts, so the discourse is not fluid.	There are a lot of pronunciatio n problems (+4). A lot of pauses, which show doubt, so the discourse is not fluid.	The pronunciation problems affect comprehension . Long pauses, doubts, and the discourse is not fluid.
Compre-hension and produc-tion	Successful completion of the task. The grammar is not limited to basic structures. There is communicati on between	Good understanding of the purpose of the task, and appropriately use of the vocabulary/grammar . There is communication between the members of the group and the audience.	Understands the purpose of the task, although the project is not complete or logic. The members of the group do not participate equally in	The group does not show understanding of the purpose of the project and the work is not complete or logic. The information presented is too basic or inexistent.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

	the members of the group.		the completion of the task.	
Originality	Unique project, which stands out from the rest. It shows creativity, freshness and uniqueness.	The project meets the requirements, but it's not unique or original. It has components that are repeated in all the other projects.	The project is not well-organized. It is very repetitive.	Very little preparation shown by the students.

Table 4-34: ***The students' grading rubric English version***

RUBRIC TO EVALUATE PRESENTATIONS (students)

Name: _____

Presentation number: _____

1. Between 1 and 10, being 10 the highest score, how would you evaluate the originality of this group's video?

   1 2 3 4 5 6 7 8 9 10

2. Between 1 and 10, being 10 the highest score, how would you evaluate the clarity of this group's presentation?

   1 2 3 4 5 6 7 8 9 10

3. Between 1 and 10, being 10 the highest score, how would you evaluate this group's preparation?

   1 2 3 4 5 6 7 8 9 10

Do you have any feedback for the group?

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Contributor's Questionnaire Responses

Section One: Test Planning and Writing

Why did you write the test? What were the purposes of the test?
This test was created to evaluate students' progress in a conversation class. My students are used to either memorizing what they have to say, or to write a lot in their exams. However, no other professor had tried to measure their real speaking performance, and I tried to change that with my final project.

The evaluation of the course was divided between a midterm oral exam (a personal interview with me to help me establish a "starting point" from which to evaluate students' progress), attendance and participation, weekly homework (including writing, reading, listening and speaking activities), and this final project. Although students' final evaluation followed those four aspects, this final project accounted for 40% of their final grade.

The main idea for me was to create a test that would allow me to have a real assessment of their progress in their speaking skills. Most of the students had memorized parts of the midterm oral exam, and so I was not able to evaluate their real production. I used this test to measure their progress and their speaking skills, and each student received a final report including information regarding his/her individual progress, his/her participation in the group, and the feedback students gave to the group regarding how original the project was and how well they had prepared to complete this assignment. See Table 4-35.

Table 4-35: ***The final project evaluation form***

COMMENTS ON YOUR FINAL PROJECT AND GRADE

Notes and commentaries:

1. Evaluation and average grade give to your group by your classmates (10% of final grade)	
• Between 1 and 10, being 10 the highest score, how would you evaluate the originality of this group's video?	/ 10
• Between 1 and 10, being 10 the highest score, how would you evaluate the clarity of this group's presentation?	/ 10
• Between 1 and 10, being 10 the highest score, how would you evaluate this group's preparation?	/ 10

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Feedback given to you by your classmates:	
AVERAGE GRADE:	/ 10
2. Evaluation and average grade given to you by the professor regarding your group work: It includes the originality of the video and the division of the work between the members of the group (40% of final grade)	
Feedback:	/ 10
3. Evaluation and average grade given to you by the professor regarding your individual work (50% of final grade)	
Feedback:	/ 10
FINAL GRADE:	/ 10

How did you decide how many subtests to write? Was there some correspondence between a subtest and a course objective, or was there some other reason to create the subtests you did?

In my final project, there is not really something like a "subtest." There are two parts to this test. The first part is a video that students would need to record and prepare in pairs/small groups, and the second part represents their presentation of their video, putting in the voice "in-off" in front of everybody in the classroom and answering my final question (for which they could not prepare).

My decision on making this division was to give students time to prepare the video, working in small groups to really think about the topic and discuss among them the vocabulary they would need to use, the structures they could use, and the main ideas. Although not directly stated, this was my way to make sure they were preparing for the final presentation, without directly memorizing what they would need to say later on. The second part, which is also the one that obtains the highest credit, is the part in which they need to speak in public and that they cannot fully prepare (their personal answer to my final question).

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you write as many items or test tasks as you needed, or did you write more than you needed, then pare the number down? How did you decide which items or test tasks to keep? How did you decide which items or test tasks to discard?

I honestly started by including three parts of the video, the first of which included a writing section. I was going to ask them to write down their answer to the specific question they had and for which they had to give an answer using the video/in-class presentation. However, after thinking about it, I realized that this task was not really related to the purpose of the class (improve their conversation skills) and so I decided to not include it. Rather, I encouraged students to really work on preparing the video for the presentations, and, indirectly, pushing them to prepare the vocabulary/grammar they would need for the presentation.

Did you have one version of your test, or did you create a second equivalent version?

I created several versions (with different topics, including the writing section versus deleting it, etc.), but I ended up just administering the one I am presenting together with this questionnaire. All students completed the same final project. In the future, I hope to be able to use the different versions originally prepared in future classes.

Were you concerned at how long the test would take to administer?

Not so much about how long it would take to administer, but how much time it would require for students to complete this project. The hardest part for them seemed to be to come up with a topic they would like to discuss and to prepare the video, but students said once they had a topic, they were done with it (they prepared the scripts and recorded the video) within a couple of days.

Were you concerned at how long the test would take to score?

Honestly, I did not think about this part when I was preparing the test, something that I regretted later on, when I started to evaluate the final projects. The quality of the videos was higher than I had anticipated, and students had worked really hard on their part of the presentation. In fact, it was the way in which they answered my final question that made the difference when evaluating the work of students.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Were you concerned how you might use the test items themselves for learner feedback?
From the beginning I knew I would be able to use my final question as a clear assessment of their conversation skills' progress. I would be able to determine whether they could understand my question (each group/student received a different question) and how they would be able to address it with respect to their level of proficiency and in relation to their "starting point" (the midterm oral exam).

Did you consider having your students take your test on a computer? Why or why not? If not, what were your concerns?
No. While there may be options to do it that way, I still prefer the in-class presentation method. I think the in-class presentation puts everybody at the same level in terms of external factors that could affect their performance. For example, some students feel more comfortable in front of a camera than others, thus those who are not so familiar/comfortable with talking in front of a computer could be at a disadvantage, not because they are less proficient in Spanish, but because they feel uncomfortable in that given situation. To avoid these external factors from introducing bias into my students' performances (and grades), I preferred asking everybody to do the same procedure.

Did you consider making the test an open book test?
With this type of test, having the book with them would not be particularly helpful, so I did not even consider the possibility of letting them bring their books.

Did you consider allowing learners to use additional sources such as dictionaries, or their notes, while taking the test?
I allowed them to bring notes with them. The notes helped them during the presentation, but they were asked not to read them, but just have them as support. The notes did not help during the final part of the presentation (my personal question).

Did you plan to allow learners to re-take a test for improvement? The same test, or a different test?
No. I gave them around two months to prepare for this final project, and they all knew that it was the most important part of their final evaluation and that they needed to prepare as carefully as possible.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

What sources did you draw from for your test items?
In 2017, I participated in the Fundamentals of Project-Based Language Learning Online Institute, offered by the National Foreign Language Resource Center (NFLRC) at the University of Hawaii at Manoa (NFLRC, 2019). In this online institute, we learned about newer ways of approaching the teaching of foreign languages and content topics in a foreign language, in a way that is more meaningful for students. For example, we learned about how using topics interesting for students would make them feel more engaged with the learning process itself. Thus, the final product would be closer to students' actual level of proficiency because they did their best with something they like, rather than repeating what they had previously memorized.

Based on what I learned in the institute and some of the ideas that were discussed during the online lectures, I created this final project. The topic of education, for example, was discussed in class, but specifically asking them my personal question about what the university means to them was my idea. I thought about this specific question because it is a topic that is important for them right now. For instance, we do not have a library, as it is under renovation, and most of the buildings do not have an elevator and students with disabilities cannot attend those lectures, etc. I thought it would be interesting for them to look back at their time at the university and consider whether it was positive, negative, and what aspects they consider important.

When evaluating them I did try to follow some of the guidelines or descriptors proposed by ACTFL (American Council on the Teaching of Foreign Languages, 2012a). That is why I created a task that would engage them in free production (at least the last part of the presentation), while motivating them to carefully prepare for the presentation.

Which test item formats do you prefer to use?
I normally ask students to complete tests that include short answers, fill-in-the-blank, and matching tasks. However, none of these tasks made sense to evaluate students' conversation abilities. That is why I decided to change the format of the final project in this case. I would describe my final question as a short answer type of task, targeting free production.

What types of learner knowledge do you believe you are capturing in your test?
My aim was to try to capture the level of grammar and vocabulary that Spanish learners are expected to have at an intermediate level of proficiency. Again, my aim was to capture the knowledge they have and the use they can make of it in a context in which they cannot fully prepare their exam (like, for example, memorizing sentences and just copying them when completing the exam).

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

What learner skills do you believe you are capturing in your test?
My aim is to capture learners' speaking and listening skills (primarily speaking skills). I think this test managed to capture this specific set of skills, mostly with the last question, which targets free comprehension and production.

If you wrote a performance test (where learners have to converse or present a topic, or write something above the sentence level), how did you get the ideas for what task learners had to do for the test?
I got my ideas from the real experiences that my students are having while being students at the university, and the ideas that were discussed in an online course, the Fundamentals of Project-Based Language Learning Online Institute, described above.

How did you get ideas on how to score learners' performances (the scoring criteria)?
This was one of the hardest parts, when I was creating this final project. I looked for ideas online, as well as ideas from the Online Institute mentioned earlier and the courses where I worked as a teaching assistant in the U.S.A. I wanted to create something that was unique and principled, but I also wanted to make sure that I was being fair and evaluating students' final project in a way that really captured the work done and the progress made during the semester.

If you used points on a scale for scoring learners' performances, how did you decide on how many points on a scale to use?
I only used a scale in the feedback that students provided to their classmates (Table 4-35). I just used a 1 to 10 scale, because I felt it was easier for students to relate this scale with the grades they normally obtain in class, and also because it was easier to calculate the final percentage obtained in this specific aspect of the evaluation of the final project.

Were you concerned about whether you could get a colleague to help you score learners' task performances?
Consider: Did you have ideas about how to ensure score consistency, such as having two scorers (you and a colleague), or scoring learners' performances by yourself on two different occasions?
Asking a colleague was not an option, as that is not the way in which things are done in South Korea. My colleagues experience a great amount of work towards the end of the semester, so I did not want to add them extra work by helping me evaluate my students. I knew I had to evaluate everything by

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

myself, so I decided to evaluate the final projects myself on two different occasions.

On the first occasion, I gave them a score shortly after they had finished their presentation and I had been taking notes on their performance. Once I was back at home, I evaluated everybody's performance a second time, this time randomizing the order in which I reviewed the presentations/videos and I looked at my own notes.

Did you plan to give the scoring criteria to the learners for future self- or peer-assessment? Did you make another, perhaps simpler or shorter, version of the scoring criteria for learners to use?

From the very first moment, I gave them the scoring criteria in Spanish, including the students' evaluation form (Table 4-32), together with the instructions on how to complete the final project. I wanted to make sure they would understand how I was going to be evaluating their work even before they started working on it. Understanding how important the presentation was would make them more conscious regarding that part of their project and more motivated to do their best. To make sure they understood the instructions and rubrics, we went over them together during part of one of the class sessions.

Did you seek help from a peer to clarify what your test items or tasks were measuring?

Consider: Did you state at any point what you wanted to measure in your test?

Did you ask another teacher to compare your test items with what you said you wanted to measure?

Did you make any changes to your test or items as a result of your colleague's feedback?

I asked for feedback from former colleagues who have used tasks like the one I used. I wanted to make sure that 1.) My task was indeed appropriate for an intermediate level, 2.) The amount of work was fair as a final project targeting conversation skills, and 3.) That the wording used in the instructions was clear and concise. I adjusted the wording of my instructions, and I decided to remove the writing section of the project after hearing their feedback.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you compare your test to the lessons that learners had?
I did, to make sure I was not asking them to do something beyond their current level of proficiency.

Did you compare your test to the textbook or other materials learners used?
I did, to make sure I was not asking them to do something beyond their current level of proficiency.

Adapting existing tests

Are you required to use specific tests in your program?
No, I have complete freedom in the decision of what type of tests to use. However, I am expected to use a test that targets the main content covered in the class (e.g., speaking test for the conversation class) and keep some sort of record in case the university wants me to explain my decision on certain grades.

Did you inherit your test or parts of your test?
No, I created everything by myself, although it is true that I gathered ideas from the examples we were given at the National Foreign Language Resource Center Institute.

Section Two: Test Administering and Scoring

Were you concerned about test security? What did you do to ensure test security?
I was really concerned about test security. One of my problems is that I could not make sure that students were not copying their ideas from other sources, for example, by reading/watching what other students had done with similar projects. That is why I decided to include the last question, which I personalized for each student/group, based on what they said in their presentation. This was my way of trying to avoid any type of cheating, as they could not prepare for it.

Were your tests photocopied, or did learners see your test items or test tasks another way, such as on the blackboard?
I uploaded the instructions on how to complete the final project two months before the deadline using the university's blackboard system. In Korea, it is called E-Class, and it is the site, within the university's website, used to interact with students and share the materials/announcements/etc. with them. All the students had access to the instructions at the same time and all of

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

them could access them as many times as needed and even download them to their own devices if they wanted to.

How did you deal with learners who missed the test, or who were late for the test?
They knew this was their final project/exam and that, if they missed it, they would get a 0 in the presentation part. As this was the part of the evaluation that counted the most for their final grade, so nobody missed it or was late.

How did you prepare learners to take the test? Were there any test items or test procedures that were new to them?
Everything was new for them, as I am the only professor in my university who has replaced the final interview with final projects. To make sure everybody understood the main idea of the final project, I used part of one of the lessons to go over the instructions with them.

In the instructions themselves, I included several examples of videos they could watch to get ideas on how to develop this final project. And, in class, I gave an example about how a presentation, including the voice-over part of the project, should look and sound like.

Did you pilot your test? Do a trial run?
I did not, although this is an aspect I will take into consideration when preparing future final projects like the one I implemented, and am sharing in this book chapter.

Did you write any of the test in the learners' first language? Why?
No, because my Korean level is not advanced enough to write anything like what I had in the instructions. Moreover, their level of Spanish was advanced enough for them to understand the instructions, although I used some English words when I was describing the task in class for them, just to make sure everybody was following me.

Was your test administered on a computer?
No. While students used computers to prepare their presentation (e.g., editing the videos, etc.) and deliver it in front of the class, I don't think computers alone are a good way of conducting a type of test that requires learners to interact with the audience and answer questions. I think it could be doable in circumstances in which the students cannot be present in-class. However, I still favour the option of doing presentations in person.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

For classroom tests: How did you accomplish scoring learners' tests? Consider: Did you score learners' tests twice for accuracy? Did you hide learners' names as you scored? Did you go back and change your marks on previously scored tests in response to problems you found while scoring tests later in the process?
As I mentioned before, I knew I had to evaluate everything by myself, so I decided to evaluate the final projects myself on two different occasions.

On the first occasion, I gave them a score shortly after they had finished their presentation and I had been taking notes on their performance. Once I was back at home, I evaluated everybody's performance a second time, this time randomizing the order in which I reviewed the presentations/videos and I looked at my own notes. Normally, both evaluations were quite similar, although in some cases I realized I was being too harsh with some students (taking off points for mistakes they were making that were beyond the expectations of the course). Thus, I think this double process of evaluating them allowed me to be equally fair with everybody.

It was impossible to hide the identity of the students' whose work I was evaluating at each point, as I could see them in the videos and hear their voices in the recording I did of their presentations. However, as I mentioned, I tried to counterbalance the other in which I evaluated them to make sure no personal reasons (e.g., being tired) affected negatively any of the students.

Did you put learners' responses to items into a spreadsheet for further analysis? Did that process help you catch scoring accuracy problems? Or problems with bias?
I did. I created a spreadsheet that was automatically calculating their final grade (considering the different percentages that each part of the project was worth it, such as 10% for their classmates' feedback). In this spreadsheet I included the grade for each one of the portions of the final project, as well as my notes on each specific aspect (e.g., my notes on their individual and group participation). I used these notes and grades to complete the form that I sent to each individual student explaining them their final grade.

Personally, I thought that this process made it easier for me to make sure I was being equally fair with everybody. This is how I realized in some cases I had been a little bit too harsh with some students.

Did you ask the students themselves to score their own test? Or a classmate's test?
Actually, I did. As it can be seen in the test I attached to this questionnaire, I asked all the students to individually provide feedback to the other groups using the students' scoring rubric in Spanish (Table 4-32). I wanted to make

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

sure students were engaged and paying attention to other students' participation and to judge these students' efforts. I did not ask them to evaluate their classmates' grammar/vocabulary (that was my job), but the effort put into their final project.

For performance tests: How did you accomplish scoring learners' performances?
As mentioned before, I knew I had to evaluate everything by myself, so I decided to evaluate the final projects myself on two different occasions. Normally, both evaluations were quite similar, although in some cases I realized I was being too harsh with some students (taking off points for mistakes they were making that were beyond the expectations of the course).

Do you think your test was reliable? What did you do to check?
This is a hard question to answer. I think my test was reliable, as the final grades obtained correlated with other measures of students' proficiency I collected during the semester (weekly homework assignments and midterm scores). However, apart from this correlation, I do not have any other "proof" of its reliability, other than my own judgement.

Reporting scores

Did you report the scores to learners? If so, how did you report test scores to learners?
I used this test to measure learners' progress and their speaking skills, and each student received the final project evaluation form (Table 4-3●), including information regarding his/her individual progress, his/her participation in the group, the feedback students gave to the group regarding how original the project was, and how well they had prepared to complete this assignment.

What was your goal in reporting the test scores to learners?
My goal was to make sure students understood their grade and that they could use my comments to help them further improve their speaking skills. I gave them additional feedback on general aspects, such as "you need to rely a little bit less on your notes when presenting" and on specific aspects of their grammar/vocabulary.

Did you teach learners how to interpret their test scores?
I did not teach them how to interpret their test scores, because I thought the form was quite straightforward and easy to interpret.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Did you report the scores to anyone else?
No.

Did you spend time explaining scores, or answering learners' questions about scores in or out of class?
I did not spend time in class explaining scores or answering learners' questions, mostly because the presentations were done the last day of classes, and we did not meet again in the classroom setting. However, I did spend some time out of class answering their questions, in person or by email.

Did you report peer-assessment scores or self-assessment scores on the test?
I included peer-assessment scores (worth 10% of their final grade) in their forms (student evaluation form, Table 4-32, and final project evaluation form, Table 4-35). I wanted students to also understand how their project was perceived by their peers.

How quickly did you report scores to learners? Was speed a priority?
Within a week after their presentations. As presentations were done during the last week of the semester, and before they completed their final exams for other courses, I had to submit these forms to the students relatively quickly to make sure we still had time to go over their grades together in case they had any doubts/questions. Another source of time pressure was I had to submit the final grades to the administration of my university.

Section Three: Using Test Scores

Cut scores

How did you decide which scores were passing or failing scores (cut scores)?
How did you decide which scores meant a specific grade on a test?
Consider: Did a language use framework such as CEFR or other standards help you determine cut scores?
Did your institution stipulate cut scores?
In order to establish the cut scores, I followed two criteria: The evaluation system used in South Korea and the standards established by the Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001).

South Korea follows a relative grading, which is a government imposed grading curve. While it depends on schools, this grading curve demands that at least 30 percent of students are destined to receive a grade of C+ or lower.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

So, even when students deserve to get a higher level, the grading system at the university does not allow professors to give more than a certain percentage of As, of Bs, and of Cs.

The class in which I used this test is supposed to start with students with an A2 level and prepare them to take (and pass) the B1 level (as, for example, in the DELE exams). I followed the criteria examiners from the Instituto Cervantes use when evaluating the candidates taking the DELE B1 exam (Instituto Cervantes, 2019), always taking into account the grading curve I am supposed to use at all times.

How did you use learners' scores from this test?
Consider: Were the scores for your use only?
Did learners' test scores have any positive or negative consequences for you, in terms of your institution?
The learners' scores from this test were just used by me, and there were no positive or negative consequences for me. Nobody else has had access to my scores (or the results of the test themselves) other than me.

What was the role of the test score in determining learners' grades?
Consider: How much weight did you give your test? How did you decide?
Were other measures used to decide learners' grades, besides your test?
What was the relationship of the other measures to your test?
Did your test capture some knowledge, skill, or ability the other measures did not capture?
This was the final project for the class (replacing the traditional final exam), worth 40% of students' final grade. The rest of the scoring criteria was divided between weekly assignments (20%), midterm oral exam (30%), and participation (10%).

From the beginning, I knew this final project should be the aspect to receive more credit, because it would require quite a bit of preparation/work from the students, and I wanted to compensate that effort. Moreover, it was going to be the piece of evidence that would provide me with a clear account on students' real speaking improvement, as they would not be able to prepare everything beforehand and they would need to produce free speech.

Reporting scores

How did you report scores to learners? Was timeliness of concern to you?
I sent individual reports to each one of them, within a week after their presentations. As presentations were done during the last week of the semester, and before the completed their final exams (for other courses) and

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

I had to submit the final grades to the administration of my university, I had to submit these forms to the students relatively quickly to make sure we still had time to go over their grades together in case they had any doubts/questions.

Did you hand the test back to learners?
No. Students only received back the individual reports/forms with my scores and feedback.

Did you offer feedback to individual learners in addition to their test scores? Written? Orally? In or out of class?
In the individual reports, I provided students with feedback in addition to their test scores. My goal was to make sure students understood their grade and that they could use my comments to help them further improve their speaking skills.

Did you teach learners to interpret their scores?
I did not teach them how to interpret their test scores, because I thought the form was quite straightforward and easy to interpret.

Using test scores

Do you feel you learned what you needed to learn from the test results about what learners could and could not do, or what they knew or did not know?
I do. This was the second semester I was teaching in South Korea and I had already learned that students here are really good at memorizing things and preparing for written tests. However, I wanted to make sure I could target free production, something that they could not explicitly prepare for. I think this test, while I am sure it could be improved, managed to get me the type of information I was looking forward from my students' progress and their speaking/listening skills.

Did your test change how learners studied?
I think it did, and this is one of the more interesting aspects of my final project. South Korean students are used to memorizing large amounts of information and just repeat that information in their tests. However, I wanted to create a task that would make them think critically while making real use of the language. Again, this final project may not be perfect, and I am sure there is plenty of room for improvement. However, I think that it achieved the purpose of making them use the language to express their own

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

thoughts rather than just repeat sentences they had previously memorized from their grammar books.

I am not sure whether I can claim that they had changed their studying habits based on the test results. However, students came to me at the end of the semester and told me in person how much they liked my class (including this final project), because they were able to really use the language for the first time in their studies.

Did you spend time going over the test in class?
No. The final project was presented during the last lecture of the semester, so we did not have any more time together to go over the test results in subsequent classes. However, it would be ideal to have some extra time at the end of the semester to give feedback on how to improve their presentations/final projects. This is something I will try to incorporate in future classes.

Did learners ask you about the test itself (not the test scores) outside of class? If so, what did they want to talk to you about?
They mostly wanted to make sure they had understood the instructions or whether they could add some extra examples/details that, even not explicitly stated in the instructions, that could help them better prepare for their final projects.

Did learners' test scores change your teaching?
Consider: Did you change your teaching for future courses based on test results?
Not for the students who completed this final project, unfortunately. However, I think the results of this experience have changed how I approach the teaching of my conversation classes since then. After this "experiment" last Spring semester, I have included many more group activities in my classroom, most of which require students to solve some sort of problem in groups.

If you could turn back time, what would you change about your test? What would you change about your test administration?
Consider: Did learners give you feedback on the test? Did they think the test was fair, or helpful?

I think I would give myself more time to go over the results of the test with the students and give feedback that could serve not only to the individual, but also to the whole class. I used the results of the test to give individual feedback to students.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

However, I feel I could have given some more general feedback to the classroom, which could have helped students better prepare for future presentations. For example, I think I would do two projects, some sort of midterm project, which could serve as a baseline to determine students' "starting point" and the final project. Then, students could make sure to put into practice the feedback I would give them after the midterm when preparing the final project.

Section Four: Evaluating and Reviewing your Answers

To what extent do you think you've described recurrent patterns in your work with tests?

I have talked all the time about the importance of the final report I submitted with the feedback and the final scores, and the importance of using this feedback, hopefully, in a more general way, so that students can implement the feedback in future classes/presentations. I think this is something recurrent in my work with tests as it is very important for me to make sure students understand their scores and have some sort of information for them to know how to improve in future classes/in their Spanish.

I have also emphasized a couple of the times in my responses to the questionnaire the importance of the training I received in the Fundamentals of Project-Based Language Learning Online Institute, offered by the National Foreign Language Resource Center (NFLRC) at the University of Hawaii at Manoa. While I modified the examples we discussed in the institute to match my students' needs, it is evident to me that learning more about new ways of testing students was critical for me to change the way in which I approach my teaching. The training is also something recurrent in my work, as I always try to keep up-to-date with the latest pedagogical approaches.

Finally, I think have also emphasized several times how important it was for me to create a test that would help me target free production. My students were used to conversation classes in which they had to complete a writing test as their final exam. However, this was something that needed to be changed. While not perfectly, I am sure, this test is the beginning to try to set up a curriculum that emphasizes the specific skills that the classroom is meant to teach. Understanding that a test should really target the language skill practiced in the classroom is another aspect that is recurrent in my work.

To what extent is your test here an innovation, or something new, for you?

I am honestly proud of this final project, and this is the reason why I decided to submit it for publication as a book chapter. While I am completely aware

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

of the fact that it could be improved in many ways, this was the starting point for me to change my teaching/evaluation habits.

I used to follow traditional ways to create and administer tests (e.g., oral interviews or written reports). However, I always felt that I was not fully capturing the real level of proficiency of my students, and that I was limiting their possibilities to freely express themselves. I think this type of project allows the teachers to evaluate the real level of proficiency of the students, while the students are doing something meaningful for them. Since then, I am trying to implement these ideas in all my courses.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

CHAPTER FIVE

COMMUNICATIVE COMPETENCE AND LANGUAGE USE DESCRIPTION FRAMEWORKS AND SECOND LANGUAGE TESTS

GRETA GORSUCH

What this Chapter is About

This chapter relates communicative competence, and second language use description frameworks such as the ACTFL *Guidelines* (ACTFL, 2012a) and *CEFR* (*Common European Framework of Reference*; Council of Europe, 2001; 2018) materials to classroom tests. There are both theoretical and practical reasons for doing so. In Chapter Two, a framework and a resulting questionnaire were proposed to probe actual tests that second language teachers make and use. The questionnaire had to be broad enough to capture all reasonable influences on teachers'/contributors' current states of teacher theory, which informs test making, scoring, and use (see the Teacher Theory model, Figure 2-1). Communicative competence, even if narrowly understood, is a salient high-level theory in second language education. There were bound to be teachers/contributors who would cite communicative competence as an active influence on their test making and use (for instance, contributor Myles Grogan, and also Maria Martinez-Garcia). And, if published reports about the ever-increasing presence of ACTFL *Guidelines* and *CEFR* in teachers' daily working lives are to be believed, the level and skill descriptors from the ACTFL *Guidelines,* or "can-do" statements from *CEFR* would be cited by teachers as inspiration or influence on their classroom testing activity (contributor Beatriz Glick mentions the ACTFL *Guidelines*, while Gisela Mayr mentions *CEFR*). The influence may occur through teacher background, current and past educational opportunities teachers have taken, and teachers' institutional

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

contexts, if their schools use the ACTFL *Guidelines* or *CEFR* for course planning or for accountability through testing (see Figure 2-1). Aside from the need to include communicative competence and second language use description frameworks in the questionnaire, any current resource on classroom tests such as *Tests that Second Language Teachers Make and Use* ought to explore, if briefly, these salient middle-level theories and their applied uses (see Table 2-1). Knowing and using theories to design and use tests is an important cornerstone for validating a test (Chapter Six).

This chapter presents definitions of key concepts, including the High Middle Low Theory (HML) model, communicative competence, proficiency, language use description frameworks, and target language use tasks (TLU). There is also be a brief section on multiple literacies, which is an incipient middle-level theory conjoining communicative competence and critical thinking. Where possible there are applied examples of the concepts from second language teaching and testing, including contributors' tests and commentaries. The examples are offered as a way to explore possible resources they offer teachers/testers to make and use classroom tests. There is a chapter summary, and further reading suggestions for applications of communicative competence and language second language use frameworks to testing projects.

Definitions of Key Concepts

High Middle Low Model

While the HML model is briefly touched on in Chapter Two, it appears here in extended form with examples and applications. The HML model is defined first because the model may assist in giving context and order to the many theories used in second language education, particularly as they relate to teaching and testing. The definition is: The HML model is a model of theories that identifies three types of theories that teachers/testers may recognize and make use of in their work. These are high-level theory, middle-level theory, and low-level theory (Gorsuch & Griffee, 2018).

Teachers do not have an easy relationship with theory (Beretta, 1991; Clarke, 1994), yet theories are used by all human actors to carry out their daily work and lives (Griffee & Gorsuch, 1999). Griffee's High Middle Low (HML) model (2012b) was an attempt to account for different kinds of theory in apparent use in second language education. At the time, Griffee wanted to isolate and study local, or "low" teacher theory. He was working with novice second language teachers, and he wanted to account for how the teachers formed over time the ability to plan and carry out their teaching.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

From this process, the HML model emerged. It seemed that classroom teachers, aside from the novice teachers he was working with, tapped into some form of private personal theory to plan lessons and teach. Their actions were internally coherent and ordered, apparently, in response to in-depth knowledge of a specific context, in other words, classrooms of learners. This theory seemed localized, private, efficient, and purposeful, and was termed "low-level theory" or teacher theory (Figure 2-1). Yet it was also clear that different actors in second language education also tapped into theories to do research, plan instruction, and write tests. These theories seemed broader in scope and designed for use in multiple contexts and across institutions. Further, the theories came from published, public sources. This type of theory was termed "middle-level theory" or "domain theory," in that this theory type spoke to specific areas or domains, such as listening comprehension, or second language learning. And, a few theories were posited that had broad and profound effects on second language education as a field. This type was termed "high-level theory" or "grand theory." The terms "high" or "middle" or "low" do not refer to the quality or usefulness of a theory, but rather to the particularity of the theory, or its range of its use.

High-level theory examples. Examples of high-level theories in second language education are communicative competence, and language proficiency (Bachman & Palmer, 1996; Celce-Murcia, Domyei, & Thurrell, 1995; Fulcher, 1998; Harsch, 2014). A third example more specific to language testing specialists would be test validity models, such as Kunnan's (1998) interpretation of Messick (1989). High-level theories are public, and widely discussed and cited at both conferences and in publications. High-level theories in second language education have profound and lasting effects, sometimes unintended, on many levels of the educational enterprise, including textbook design, classroom instruction, and teacher education. High-level theories establish fundamental and overt changes in theory and practice in the field, and answer the question of what reality *should* be. In sum, high-level theories change how second language professionals view their working universe, just as a theory of plate tectonics has changed scientific, and even popular, conceptions of earthquakes and volcanic activity (Winchester, 2004).

Middle-level theory examples. Examples of middle-level theories come from fields such as applied linguistics, communication studies, education, and psychology. Middle-level theories deal with specific domains of interest such as how listening comprehension takes place (e.g., Rost, 2002), how learners plan doing tasks based on experience (Self-Efficacy Theory, e.g., Bandura, 1997; Siegle, 2000), and how learners learn a second language (second language acquisition theories, e.g., Shehadeh,

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

2002; Van den Branden, 1997). Middle-level theories are used to motivate research agendas, among other things, and may or may not be intended for classroom applications.

A notable example of middle-level theory that is intended for both testing and classroom applications is language use description frameworks, or models, such as the ACTFL *Guidelines* (American Council on the Teaching of Foreign Languages, 2012a) and the Common European Framework of Reference (*CEFR*; Council of Europe, 2001). Is it suggested here that the ACTFL *Guidelines* and *CEFR* are conceptually akin to high-level theories of proficiency and communicative competence, respectively. As will be seen in the following pages, terms such as proficiency and communicative competence have been used interchangeably in the field. And, the ACTFL *Guidelines*, for a variety of reasons, has become functionally aligned with a curricular movement in second language education called "the proficiency movement" (see Ringvald, 2006 for an example of this).

Second language use description frameworks were intended as a means of promoting accountability and a common understanding of second language ability across institutions (Phillips & Abbott, 2011) and across borders, in the case of *CEFR* (Council of Europe, 2001; Fulcher, 2004). Authors of the two frameworks, over the years, developed intuitive descriptions of second language use in broadly stated academic, cultural, and social contexts. They then put the descriptions on a scale. Some scholars describe their widespread influence (Alderson, 2006; Liskin-Gasparro, 2003; Phillips & Abbott, 2011). This may give the impression that the ACTFL *Guidelines* and, more particularly, *CEFR*, approach being high-level theories such as communicative competence. Yet these frameworks, or models, have specific domains of interest, namely testing, curriculum planning and design processes, and accountability within and across institutions (Hatasa & Watanabe, 2017). And while *CEFR* motivates useful, applied research agendas which consider the effects of context on language use, communicative competence still forms the landscape *CEFR* inhabits.

Low-level theory examples. Examples of low-level theories, or teacher theory, are answers or action plans teachers formulate to questions like "What works for me and why" or "How my students learn" or "What I put on my mid-term exam" or "What I think about students' test scores" and "How I make future plans according to the scores" (Gorsuch & Griffee, 2018, p. 79; see also Table 3-1 in Chapter Three). Low-level theories are used to solve the "problem" of teaching, and as argued in Chapter Three, classroom testing. While teachers' theories may be informed by middle-

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

level and high-level theories, their functions play out in specific classroom contexts and thus are mediated by low-level teacher theory (Figure 2-1).

Conclusion. To conclude, the HML model is a model of theories that identifies three types of theories: High-level, Middle-level, and Low-level. Theories within each type have worth, and differ in their scope and functions. Thinking about theories this way brings teachers/testers closer to theories and the benefits they can bring to day-to-day practice.

Communicative Competence

Communicative competence is a high-level theory that seeks to explain language use as cognitive and social events. A high-level theory in second language education, communicative competence has had profound influence on the working landscapes of teachers/testers, second language researchers, schools, publishers, language testing specialists, and testing companies.

Historically, different communicative competence models have emerged (see Celce-Murcia et al, 1995; and Fulcher & Davidson, 2007 for accounts). But generally, communicative competence characterizes language use as an interaction of an individual's language knowledge, his or her metacognition (the ability to plan and monitor), and the characteristics of the language use situation. As an example, Bachman and Palmer, testing specialists, (1996, pp. 61-62) note: "in using language to express, interpret, and negotiate intended meanings, language users create discourse." Thus, language use is "multiple interactions" among the "various individual characteristics of language users" and "the characteristics of the language use…situation" (p. 62). Fulcher & Davidson (2007, p. 37), also testing specialists, note that any model of language ability has three dimensions: A model of knowledge ("what it means to know a language"), a model of performance ("underlying factors related to the ability to use language"), and actual language use ("how we understand specific instances of language use").

Second language testers have focused on communicative competence as a way to understand two things: 1.) What a leamer's score on a test means, and 2.) Whether a leamer's score on a test of some aspect of second language use can be generalized to the leamer's language use in another language use situation. In terms of making tests, communicative competence offers resources to teachers/testers by greatly widening possibilities for what learners are asked to do with language on tests. Instead of responding to multiple choice questions or answering discrete items on a paper and ink test, learners may instead be asked to interview each other (see Myles

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Grogan's contribution) or write an argument in response to a reading passage (see Juliana Jandre and Vander Viana's contribution) or order an item online (see Irina Drigalenko's contribution). Communicative competence also offers ways to score learners' language use, such as focusing on learners' appropriateness in making requests, or their handling of content during group work (see Annis Shaver's contribution).

Other scholars have been interested in communicative competence as a means of course design and language teaching. Celce-Murcia et al noted that available knowledge of second language use needs to be made available, "that is consumable for classroom practice" (1995, p. 29). Pursuing these goals has produced categorizations of language knowledge, such as Celce-Murcia et al's model which includes discourse competence, linguistic competence, actional competence, sociocultural competence, and strategic competence. Arguably, these are valuable resources for teachers/testers for test design. See Table 5-1.

Table 5-1: ***Second language knowledge in Celce-Murcia et al's (1995) model of Communicative Competence***

Language knowledge component	Definition and examples
Discourse competence	p. 13 "Discourse competence concerns the selection, sequencing, and arrangement of words, structures, sentences and utterances to achieve a unified spoken or written text" Cohesion (parallel structure, conjunction, etc.) Deixis (spatial, *here, there*, etc.) Coherence (thematization and staging, management of old and new information, etc.) Genre/Generic structure (narrative, interview, etc.) Conversational structure (topic establishment and change, etc.)
Linguistic competence	pp. 16-17 "It comprises the basic elements of communication: the sentence patterns, the constituent structure, the morphological inflections…lexical resources…phonological and orthographic systems needed to realize communication as speech or writing" In addition: Formulaic constructions, lexical phrases Syntax (word order, modifiers/intensifiers, etc.) Morphology (parts of speech, etc.) Lexicon (words, routines--word-like fixed phrases, etc.) Phonology (segmentals, suprasegmentals) Orthography (letters, spelling, phoneme-grapheme correspondences, etc.)

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Actional competence*	p. 17 "competence in conveying and understanding communicative intent" Knowledge of language functions Interpersonal exchange (greetings, leavetakings, expressing and acknowledging gratitude, etc.) Information (reporting, remembering, etc.) Opinions (approving, disapproving, etc.) Feelings (expressing and find out about love, happiness, pleasure, etc.) Suasion (suggesting, requesting, instructing, giving orders, etc.) Problems (complaining, criticizing, blaming, regretting, etc.) Future scenarios (expressing and finding out about wishes, desires, etc.) Knowledge of speech act sets Patterns of interaction that are "highly conventionalized" (p. 21) including expressing an apology, expressing responsibility, offering an explanation, etc.
Sociocultural competence	p. 23 "the speaker's knowledge of how to express messages within the overall social and cultural context of communication" Social contextual factors (participant variables: age, gender, etc.; situational variables: time, place, etc.) Stylistic appropriateness factors (politeness conventions, degrees of formality, etc.) Cultural factors (sociocultural background knowledge: living conditions, major values, awareness of major dialect differences, strategies for cross-cultural communication, etc.) Non-verbal communicative factors (body language, non-verbal turn-taking signals, use of space, non-vocal noises, etc.)
Strategic competence	p. 26 "knowledge of communication strategies and how to use them" Avoidance or reduction strategies (topic avoidance, etc.) Achievement or compensatory strategies (circumlocution, all-purpose words, literal translation from L1, etc.) Stalling or time-gaining strategies (fillers, self and other-repetition, etc.) Self-monitoring strategies (self-initiated repair, etc.) Interactional strategies (appeals for help, repetition requests, verbal and non-verbal expressions of non-understanding, rephrasing, comprehension checks, etc.)

**Note.* Authors note these descriptions are "for oral language" and that a "parallel list of specifications is needed for written language" (p. 22).

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

There is no one unitary, agreed-upon communicative competence model. In other words, communicative competence is not a unitary law or a fact, but rather has evolved over time (see Celce-Murcia et al, 1995 for a clear depiction of the early part of this conceptual evolution). It exists happily and usefully in different permutations in the way language knowledge is categorized and described (see Bachman & Palmer, 1996). See the description of Fulcher's application of Widdowson's model of communicative competence to a reading test in Further Reading at the end of this chapter.

Narrow and broad applications of communicative competence in classroom tests. Communicative competence has been construed or used in the field in both narrow terms and broad terms. In narrow terms, a common working understanding of communicative competence might be "knowing what to say and how to say it." Many college-level second language course syllabuses will include some goal statement that stands in for communicative competence, such as "The student will begin to able to express, negotiate, interpret meaning, and communicate appropriately by using language X in culture X." Unfortunately, in this narrow understanding, it is quite possible for working practitioners and course planners to exclusively deal with linguistic competence (or grammatical competence, see Bachman & Palmer, 1996; see also Barrette, 2004 in Further Reading). Teachers may have perfectly valid reasons to test grammatical competence (see for example Ferit Kilickaya's contribution, and Sakae Onoda's). But this also means that teachers/testers can simply retain traditional syllabuses and tests which emphasize syntax, vocabulary, etc. while not treating other aspects of communicative competence, such as conversation structure or knowledge of speech acts (Table 5-1; see Barrette (2004) and Fox (1993) under Further Reading). In such a conception, in terms of test design, there may not be full consideration of how different tasks (different language use situations) require different strategic competencies on the part of learners. Strategic competence itself (appeals for help, requests for repetition, etc. see Table 5-1) may not be treated as content or integrated into instruction. Learners using different features of strategic competence in interview-type performance tests may be not be rewarded if the scoring criteria, by design, do not include strategic competence. Requests for repetition may be interpreted as disfluency.

Testers/teachers have also construed communicative competence broadly in their testing projects. See, for example contributions by Meredith Stephens and Meghan Kaiser, and Maria Martinez-Garcia. A notable example found outside this book is Venema (2002), who developed a speaking course, and a parallel peer interview test with analytic scoring

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

criteria in order to give himself an opportunity "to make explicit [his] assumptions regarding oral communicative competence" (p. 1). He consciously related his scoring criteria to Canale's (1988) and Bachman and Palmer's (1996) models of communicative competence and focused specifically on "communication skills" ("pragmatic knowledge" in Bachman & Palmer's 1996 model), and "English skills" ("organizational knowledge"). Venema's scoring criteria also included a key component he called "Content" which "reflects an assessment of the relative complexity of the discourse and topic chosen" (p. 3). Students spent their time in lessons learning to talk to each other on "a narrow selection of topics." Moreover, for the test "students were fully aware of the possible topics [of the test] and were even allowed to bring notes of key vocabulary" (p. 2). Topics for the peer interview test were among topics treated in class meetings. This last scoring scale on "Content" (topics) seems small, and just good practice, like washing your hands before dinner. But it is in fact an important application of communicative competence. It bears repeating: Communicative competence posits that language use is an interaction of an individual's language knowledge, his or her metacognition (the ability to plan and monitor), and the characteristics of the language use situation. The "Content" scoring criterion, and the author's conscious alignment of the test interview topics with classroom activity topics, integrates learners' metacognition and the characteristics of the language use situation into the test itself.

Conclusion. In conclusion, communicative competence is a high-level theory that seeks to explain language use as cognitive and social events. There are multiple communicative competence models, which have had a profound influence on the second language education field, and by extension, to second language classroom testing. Any model of communicative competence must take into account interactions between a learners' language knowledge, their ability to perform language, and their conceptions of a specific language use situation and what the situation demands. These considerations greatly expand resources for how teachers/testers decide what learners are to do to use language in a test, and how learners are to be scored.

Proficiency

Second language proficiency is also a high-level theory. It can be defined as the ability to use the second language for "some future activity" (Davies, 1990, p. 20). Proficiency is, in many ways, defined by how it is tested. Davies (1990) compares achievement tests and proficiency tests, noting that an achievement test "refers back to previous learning and is concerned

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

solely with that" (p. 20) while "the *proficiency* test is also interested in what has been learnt but in a much more vague way...it exhibits no control over previous learning" (p. 20; emphasis in the original). In other words, a learner's proficiency, as measured by a proficiency test, "establishes generalizations on the basis of typical syllabuses...and is more directly related to what it attempts to predict" which is some future use of the language (p. 20).

Testing specialists have spent a good deal of effort and discussion on defining some conceptual aspect of proficiency, and then finding evidence for its conceptual structure through carefully developed proficiency tests administered to large groups of learners across institutions. Their interest is to write a test that will put second language learners from diverse educational contexts on a scale by which they can be compared to each other. For instance, Goh and Aryadoust (2010) explored listening comprehension on a commercially developed listening test using a variety of multiple choice test items ("minimal context," "propositional inferencing," etc.). They argued that the different multiple choice item types captured four different "competencies" (p. 35) that they believed, through previous study and research, comprised second language listening proficiency. Their test population was diverse, with 916 learners from 78 different countries who were studying English at a major U.S. university (p. 36). Such tests are made by testing specialists with a background in psychometrics, and are rarely attempted by working teachers/testers. Proficiency tests are used to compare learners to each other, but not for comparing learners to content they have learned in a course (an achievement test, or an ordinary classroom test; see Chapter Three). Proficiency as a high-level model offers few resources for making and using achievement tests, as by definition, it has no relationship to a specific course.

Over some decades, proficiency has been seen and discussed by applied linguists and language testing specialists as single global language factor and alternatively as multidimensional (Harsch, 2014). See, for example, Hulstijn's model that posits language proficiency as two elements: "Basic Language Cognition" and "Higher Language Cognition" (2011). Using this model, he attempted to explain why first language speakers of a language "differ enormously in the command of their L1" and at the same time why second language speakers of the language "differ in the success with which they acquire their L2" (2011, p. 230). In other words, he wanted to put both L1 and L2 users of a language, from greatly diverse national, educational, and cultural contexts, and with widely varying general intelligence, on the same proficiency scale. Once again we see that the purpose of a proficiency

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

test is to compare learners to each other, and address theoretical concerns such as Hulstijn's.

The commentary and research on proficiency described here suggests one significant difference between a proficiency model and a communicative competence model. Communicative competence is predicated on situating language ability in terms of specific language use contexts and how learners interact with it. Proficiency models do not necessarily take specific language use contexts into account. Interestingly, some scholars working with proficiency models refer to Bachman and Palmer's (1996) model of communicative competence as being a "multicomponent model" of "LP" (language proficiency)(e.g., Hulstijn, 2011, p. 236).

The proficiency movement. Proficiency is also defined by an American-based proficiency movement dating from the 1980s (Liskin-Gasparro, 2003). This movement has had effects not only on second language testing, but also on second and foreign language syllabus design and instruction (Center for Open Educational Resources on Language Learning, 2010). For a variety of reasons, the proficiency movement has been firmly linked to the ACTFL *Guidelines* (American Council on the Teaching of Foreign Languages, 2012a; see Bachman & Savignon, 1986; Lantolf & Frawley, 1988; Liskin-Gasparro, 2003). Furthermore, the ACTFL *Guidelines* have, over time taken on a professional folklore of beliefs that are not easily penetrated or discussed. In this set of beliefs, proficiency as a target has become conflated with professional progressivism and the primacy of speaking in foreign language classrooms (Liskin-Gasparro, 2003). "Proficiency" has been defined by teachers as maximizing student-to-student communication in class, and positioning grammar as "a support skill, rather than as a centerpiece of instruction" (Liskin-Gasparro, 2003, p. 484). This primacy of oral skills in this popular professional understanding of proficiency derives from the proficiency movement's early testing cornerstone, the OPI, or the Oral Proficiency Interview, which is an adaptation of the Foreign Service Interview (FSI) test dating from the mid-twentieth century (Fulcher, 1996; Savignon, 1985). The OPI is a pre-set procedure with one interviewer and one test candidate, with a set progression of questions. The OPI may not allow for other language use tasks that might more completely predict proficiency, such as "discussions, reports, …conversations," groups discussions, and games (Liskin-Gasparro, 2003, p. 2003).

The proficiency movement has also been linked in curriculum planning and syllabuses to the four skills of speaking, listening, reading, and writing, as in a Spanish course in Texas: "Spanish 1501 is a four-skills course" (available: https://www.ttu.edu/courseinfo/); and a "World Languages First

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Level Proficiency" German course in Iowa: "Understanding and speaking 'everyday German'"; reading and writing skills" (available: https://myui.uiowa.edu/my-ui/courses/details.page?id=877359&ci=147038). Whether the four skills are equally treated is an open question. See earlier commentary on the primacy of speaking skills associated with the proficiency movement.

Conclusion. To conclude, proficiency is a high-level theory that attempts to set both L1 and L2 users on the same scale for theoretical reasons. Proficiency is used to compare language users to each other across diverse language backgrounds, educational experiences, and life situations. As a result, much effort has been expended by testing specialists to create psychometrically designed proficiency tests. In second language education, proficiency is a general ability to use the second language for some undefined future activity. Proficiency tests cannot be used to test learners' achievement of course outcomes (see Chapter Three). As a result, proficiency as a high-level model may offer working teachers/testers few resources for day-to-day test writing and use. See, however, Frain (2009) in Further Reading.

Language use Description Frameworks

Language use description frameworks, such as the ACTFL *Guidelines* (American Council on the Teaching of Foreign Languages, 2012a) and the Common European Framework of Reference (*CEFR*; Council of Europe, 2001; 2018), are considered to be middle-level theories for the purposes of this book. It can be argued they are applications of the high-level theories of proficiency and communicative competence in testing and teaching. They are intended as a means of promoting accountability and a common understanding of second language ability (Center for Open Educational Resources on Language Learning, 2010). Creating a common understanding of second language ability is achieved by having prose descriptions of learners' abilities using the language, set on an intuited, continuous, linear scale. The ACTFL *Guidelines*, for example, have an overall scale with eleven points ranging from "distinguished" to "novice low" (American Council on the Teaching of Foreign Languages, 2012a). The newest version of the *Common European Framework of Reference* (Council of Europe, 2018) has a seven-point scale. See Table 5-2 for examples of the ACTFL *Guidelines* and Table 5-3 for *CEFR* scales and partial descriptors for writing.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 5-2: *ACTFL* **Guidelines** ***scale levels and partial descriptors for writing***

ACTFL scale level	Partial description for writing (ACTFL, 2012a)*
Distinguished	Writers at the Distinguished level can carry out formal writing tasks such as official correspondence, position papers, and journal articles. They can write analytically on professional, academic and societal issues…Distinguished-level writers are able to address world issues in a highly conceptualized fashion. (p. 11)
Superior	Writers at the Superior level are able to produce most kinds of formal and informal correspondence, in-depth summaries, reports, and research papers on a variety of social, academic, and professional topics. Their treatment of these issues moves beyond the concrete to the abstract. (p. 11)
Advanced high	Writers at…Advanced High…are able to write about a variety of topics with significant precision and detail. They can handle informal and formal correspondence according to appropriate conventions. They can write summaries and reports of a factual nature. They can also write extensively about topics relating to particular interests and special areas of competence, although their writing tends to emphasize the concrete aspects of such topics. (p. 12)
Advanced mid	Writers at…Advanced Mid…are able to meet a range of work and/or academic writing needs. They demonstrate the ability to narrate and describe with detail in all major time frames with good control of aspect. They are able to write straightforward summaries on topics of general interest. (p. 12)
Advanced low	Writers at…Advanced Low…are able to meet basic work and/or academic writing needs. They demonstrate the ability to narrate and describe in major time frames with some control of aspect. They are able to compose simple summaries on familiar topics. (p. 12)
Intermediate high	Writers at…Intermediate High…are able to meet all practical writing needs of the Intermediate level…they can write compositions and simple summaries related to work and/or school experiences. They can narrate and describe in different time frames when writing about everyday events and situations. (p. 13)

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Intermediate mid	Writers at…Intermediate Mid…are able to meet a number of practical writing needs. They can write short, simple communications, compositions, and requests for information in loosely connected texts about personal preferences, daily routines, common events, and other personal topics. (p. 13)
Intermediate low	Writers at…Intermediate Low…are able to meet some limited practical writing needs. They can create statements and formulate questions based on familiar material. (p. 13)
Novice high	Writers at…Novice High…are able to meet limited basic practical writing needs using lists, short messages, postcards, and simple notes. (p. 14)
Novice mid	Writers at…Novice Mid…can reproduce from memory a modest number of words and phrases in context. They can supply limited information on simple forms and documents, and other basic biographical information, such as names, numbers, and nationality. (p. 14)
Novice low	Writers…Novice Low…are able to copy or transcribe familiar words or phrases, form letters in an alphabetic system, and copy and produce isolated, basic strokes in languages that use syllabaries or characters. (p. 14)

**Note.* The ACTFL *Guidelines* writing descriptors amount to around a paragraph for each level.

Table 5-3: Common European Framework of Reference *scale levels and descriptors for overall written production*

CEFR scale level	Description for overall written production* (Council of Europe, 2018)
C2 Proficient user	Can write clear, smoothly flowing, complex texts in an appropriate and effective style and a logical structure which helps the reader to find significant points. (p. 75)
C1 Proficient user	Can write clear, well-structured texts of complex subjects, underlining the relevant salient issues, expanding and supporting points of view at some length with subsidiary points, reasons and relevant examples, and rounding off with an appropriate conclusion. Can employ the structure and conventions of a variety of written genres, varying the tone, style and register according to addressee, text type and theme. (p. 75)
B2 Independent user	Can write clear, detailed texts on a variety of subjects related to his/her field of interest, synthesising and evaluating information and arguments from a number of sources. (p. 75)

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

B1 Independent user	Can write straightforward connected texts on a range of familiar subjects within his field of interest, by linking a series of shorter discrete elements into a linear sequence. (p. 75)
A2 Basic user	Can write a series of simple phrases and sentences linked with simple connectors like 'and', 'but' and 'because'. (p. 75)
A1 Basic user	Can give information in writing about matters of personal relevance (e.g. likes and dislikes, family, pets) using simple words and basic expressions. Can write simple isolated phrases and sentences. (p. 75)
Pre-A1 Basic user	Can give basic personal information in writing (e.g. name, address, nationality), perhaps with the use of a dictionary. (p. 75)

**Note.* *CEFR* has three separate scales for writing, representing three general language use situations: Overall written production, Creative writing, and Reports and essays.

One first year Japanese-language course syllabus at a U.S. university states: "Students completing this course will be able to communicate orally at the ACTFL Novice-mid to Novice-high level or at the Common European Framework of Reference for Languages (CEFRL) level A1."

The frameworks have evolved over their lifetimes, with added details to descriptors, and added stipulations or descriptions of general types of language use situations, such as online interaction and "transactional interactive writing" using "notes, messages, and forms" in *CEFR* (Council of Europe, 2018, p. 95), and "interpersonal," "interpretive," and "presentational" "modes of communication" in the ACTFL *Guidelines* (ACTFL, 2012a, p. 7). Work on the ACTFL *Guidelines* began in 1982 (Liskin-Gasparro, 2003), giving the *Guidelines* a current lifespan of 36 years. *CEFR*'s first edition came out in 1985 (Fulcher, 2010a), emerging from the European "communicative language teaching movement" (p. 114).

On one hand, both frameworks have motivated discussion and research at international, national, and local levels, with calls for research on the ACTFL *Guidelines* appearing as early as the mid-1980s (e.g., Savignon, 1985). Alderson, Figueras, Kuijper, Nold, Takala, and Tardieu (2006, p. 3) point out the potential of *CEFR* to generate discussion at every level of educational organizations: "The *CEFR* is intended as a reference document for language curriculum and syllabus development, textbook writing, teacher training, and assessment." If this is true, then many aspects of the current state of teacher theory could be impacted (Figure 2-1). Edelenbos and Kubanek-German (2004) and Purpura (2016) credit *CEFR* with increasing teachers' testing literacy. Figueras (2012, pp. 477-478) believes

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

CEFR's "can-do" statements allow teachers "to tell each other and their clientele what they wish to help learners to achieve" and that attention gets shifted in a more positive direction to what learners can do, as opposed to what they cannot do. As noted above, the ACTFL *Guidelines* have become linked with the "proficiency movement" of language teaching (Fulcher, 1996), providing a "common terminology" for testers and curriculum workers (Fulcher, 2010a, p. 231). Kissling & O'Donnell (2015, p. 283) cite using the ACTFL *Guidelines* as actual learning materials in an oral skills Spanish course. Their learners studied the *Guidelines* descriptors to gain a better understanding of "oral proficiency" and their own progress in speaking. See Kissling & O'Donnell (2015) in the Further Reading section.

On the other hand, both the ACTFL *Guidelines* and *CEFR* have been criticized for lacking empirical evidence for their scales, and their descriptors. In other words, are the eleven points or the seven points on the two scales real? Can learners across diverse cultural and learning contexts be reliably placed on the scales based on the descriptors? Bachman and Savignon (1986) stated that the ACTFL *Guidelines* were "experientially, rather than theoretically based" (p. 381), meaning the levels and descriptors were based on the intuitions of the *Guidelines* writers (see also Fulcher, 1996; Liskin-Gasparro, 2003; McNamara, 1997). See similar accounts about *CEFR* (Fulcher, 2004, 2010a; North, 1995). Both frameworks have been criticized for being general so that teachers/testers cannot use them, theoretically or practically, to design specific test tasks or items (Alderson et al., 2006, p. 5; Fulcher, 2004; see, however, Figueras, 2012). The problem with this is that using, say, overall reading comprehension descriptors from *CEFR* (Council of Europe, 2018) to make test X and assigning learners to a level on the scale based on test X, does not mean that learners will get the same level on reading test Y, even though the same *CEFR* descriptors were used to inspire *that* test (Fulcher, 2004).

Finally, CEFR scales have been criticized for being too broad to show learner achievement in local contexts (Benigno & de Jong, 2016). Teachers/testers may have stated reasonable course outcomes, and designed an effective course to support learners to those outcomes. And, learners may have reached the outcomes, but their growth may not be reflected in a shift from one level to the next higher level. Indeed, the ACTFL *Guidelines* are not supposed to measure achievement, but rather proficiency, which is what "individuals can and cannot do with the target language regardless of the curriculum" (Center for Open Educational Resources on Language Learning, 2010). Thus, even though the ACTFL *Guidelines* and *CEFR* provide what may be common descriptions of language ability, the descriptors are abstract and general. Teachers/testers, however, must inhabit

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

the world of specifics. What test tasks or items should be used for learners to respond to? What texts for listening or speaking responses should be selected for the test? How should learners' responses be scored? And, what constitutes a good or poor score? This mismatch also leaves unanswered how language use description frameworks are actually applied to course design, and test design and use.

Conclusion. In conclusion, language use description frameworks are middle-level theories, and arguably, applications of the high-level theories of proficiency and communicative competence. They are intended as a means of promoting accountability and a common understanding of second language ability. In the case of the ACTFL *Guidelines* (American Council on the Teaching of Foreign Languages, 2012a) and *CEFR* (Council of Europe, 2001, 2018), this is accomplished by setting descriptions of language use against a scale. How general descriptions of language use can be used as resources by working teachers/testers is further addressed in the two sections below.

ACTFL *Guidelines*

ACTFL stands for the American Council on the Teaching of Foreign Languages, which is a professional organization that focuses on language teaching, learning, and testing for K-12 and college levels in the United States. In addition to a yearly conference, ACTFL's outreach activities are 1.) the ACTFL *Guidelines*, which are free (https://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012); 2.) ACTFL *Guidelines* examiner certification workshops (see: https://www.actfl.org/assessment-professional-development/professional-development-workshops) and "proficiency instruction" workshops (see: https://www.actfl.org/assessment-professional-development/professional-development-workshops/actfl-proficiency-performance-institute), which can be taken for steep fees; and 3.) multiple commercially developed proficiency tests based on separate language "skills," including the OPI (Official ACTFL Oral Proficiency Test) and the RPT (Official ACTFL Reading Proficiency Test), which again, are fee-based (see: https://www.actfl.org/assessment-professional-development/assessments-the-actfl-testing-office).

There are three documents of relevance to this discussion that are freely available on the ACTFL website: 1.) The ACTFL *Guidelines*, 2.) the *Performance Descriptors*, and 3.) the *Can-Do Statements*. For the ACTFL *Guidelines,* there are links for downloadable versions (ACTFL 2012a) in Arabic, Azerbaijani, Chinese, English, French, German, Indonesian, Japanese, Korean, Portuguese, Russian, Spanish, and Turkish. This

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

document will be referred to as the *Guidelines* in this chapter. The link is: https://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012.

On a second link there is a downloadable document, *Performance Descriptors for Language Learners* (2012b): https://www.actfl.org/publications/guidelines-and-manuals/actfl-performance-descriptors-language-learners. This is a twenty page document which is intended to help "teachers create performance tasks targeted to the appropriate performance range" (p. 3). In an example descriptor, a novice speaker in an interactive language use situation "Expresses self in conversations on very familiar topics using a variety of words, phrases, simple sentences, and questions that have been highly practiced and memorized" (p. 14). Presumably teachers/testers could then create tests, which ought to "be conducted in the same communicative manner in which the language was learned, practiced, or rehearsed" (p. 4). Yet soon thereafter in the document, the *Performance Descriptor* authors affirm that their high-level theory is proficiency, and describe performance and proficiency as "not the same" (2012b, p. 4). In other words, teachers/testers cannot assume that learners' "performances" on a test are indicative of their proficiency level on the ACTFL *Guidelines*. Rather, multiple tests (performances) are to be done by learners, and collectively the performances might "generally" correlate to a "proficiency level" (p. 4). This document will be called *Performance Descriptors* in this chapter.

On a third link are *Can-Do Statements* which are intended for learners to set their own learning goals, and for teachers "to write communication targets for curriculum, unit and lesson plans": https://www.actfl.org/publications/guidelines-and-manuals/ncssfl-actfl-can-do-statements. The authors claim that the *Can-Do Statements* can be used as "a starting point for...the creation of rubrics for performance-based grading" suggesting an application to classroom testing (ACTFL & National Council of State Supervisors for Languages, 2018a, p. 4). This collection of documents will be called *Can-Do Statements*.

The following discussion is an attempt to describe the ACTFL *Guidelines,* the *Performance Descriptors,* and the *Can-Do Statements* and relate them to each other. The following discussion assumes that teachers/testers need ways to make test tasks or test items for learners to respond to, that they then need ways to score learners' responses, and that they need to know what score comprises a passing or non-passing performance. The question is whether the three freely available documents offer meaningful resources for these testing activities.

The *Guidelines*. The *Guidelines* offers general descriptors for the eleven levels of ability they posit in four separate sections: speaking, writing,

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

listening, and reading. See Table 5-4 for a sample *Guidelines* speaking descriptor for Novice-mid, a commonly stated skill area and target for beginning college-level foreign language courses in the U.S.

Table 5-4: ***ACTFL* Guidelines *descriptors for speaking at the novice-mid level***

	Descriptor
General speaking level: Novice	Novice-level speakers can communicate short messages on highly predictable, everyday topics that affect them directly. They do so primarily through the use of isolated words and phrases that have been encountered, memorized, and recalled. Novice-level speakers may be difficult to understand even by the most sympathetic interlocutors accustomed to non-native speech.
Specific descriptor for Novice-mid	Speakers at the Novice Mid sublevel communicate minimally by using a number of isolated words and memorized phrases limited by the particular context in which the language has been learned. When responding to direct questions, they may say only two or three words at a time or give an occasional stock answer. They pause frequently as they search for simple vocabulary or attempt to recycle their own and their interlocutor's words. Novice Mid speakers may be understood with difficulty even by sympathetic interlocutors accustomed to dealing with non-natives. When called on to handle topics and perform functions associated with the Intermediate level, they frequently resort to repetition, words from their native language, or silence.

Note. Source: ACTFL *Guidelines*, 2012a, p. 9

In the *Guidelines,* the skill area of speaking includes broad suggestions as to language use situations the *Guidelines* descriptors are applicable to: "These Guidelines can be used to evaluate speech that is either Interpersonal (interactive, two-way communication) or Presentational (one-way, non-interactive)" (ACTFL, 2012a, p. 4). Language use situations for writing are seen as "Presentational (essays, reports, letters) or Interpersonal (instant messaging, e-mail communication, texting)" but also "spontaneous (immediate, unedited) or reflective (revised, edited)" (p. 10). The skill area of listening is identified by the authors as "Interpretive (non-participative, overheard) or Interpersonal (participative)" (p. 15). The skill of reading is thought to be "Interpretive (books, essays, reports, etc.) and Interpersonal "(instant messaging, texting, e-mail communication, etc.)" (p. 20). Thus, there appear to be three broad language use situations defined in the *Guidelines:* Interpersonal, Interpretive, and Presentational. This is

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

significant in that in order to write classroom tests, teachers/testers need to define a language use situation reflected in tasks, or test item format. But at the same time, they must also think about what they wish to score learners on. In other words, what should they focus on when learners respond to the task or items? In the case of speaking tasks that were presentations, would the scoring criteria be spoken fluency, vocabulary use, or the ability to use appropriate formulaic speech ("stock answers")? The speaking descriptors in Table 5-4 do not suggest scoring criteria.

Both task/test item format planning and plans for scoring must be informed by teachers'/testers' current state of teacher theory (Figure 2-1). If schools demand the use of the ACTFL *Guidelines* alone for accountability purposes (the "Institution" component in Figure 2-1), then teachers/testers need to relate the *Guidelines,* as they are stated here, to their current teacher theory. As noted, the descriptors are general, and would require a lot of extrapolation and thought for teachers/testers to design instruction, and classroom tests that reflect the instruction. For instance, the part of the descriptor on learners being "understood with difficulty" (Table 5-4) suggests a deficit view of language learners. Does it mean that learners need instruction on pronunciation?

There are also loose ends, relevant to testing, that are mentioned in the *Guidelines*, but for which there is little explanation. The *Guideline's* descriptors are intended to "describe the tasks that [speakers/writers/listeners/readers] can handle at each level, as well as the content, context, accuracy, and discourse types associated with tasks at each level" (ACTFL, 2012a, p. 4). To find what "content, context, accuracy, and discourse types" (p. 4) means, and to perhaps better specify test tasks/test item types or scoring criteria on tests, one must look in the *Performance Descriptors* (2012b, pp. 8-9).

The *Performance Descriptors*. The *Performance Descriptors* (2012b) identify functions, content, context, text types (perhaps "discourse types," ACTFL *Guidelines*, 2012a, p. 4), language control (perhaps "accuracy," ACTFL *Guidelines*, 2012a, p. 4), vocabulary, communication strategies, and cultural awareness as "language domains" (ACTFL *Performance Descriptors*, 2012b, pp. 8-9). It is not clear how these concepts were selected together as "language domains." According to one practicing Spanish teacher, functions, content, context, and text types refer to performance "parameters," which may refer to conditions of language use (tasks or test item formats). Language control, vocabulary, communication strategies, and cultural awareness refers to "how well" learners can do in a performance (scoring criteria)(J.M. Hernandez Lopez, personal communication, December 7, 2018). The *Performance Descriptors* also echo the three

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

"modes of communication" (2012b, p. 7) mentioned in the *Guidelines* with more specific examples. The examples may help with designing test tasks and test item formats that learners respond to. See Table 5-5 for definitions of the terms with examples for some of them, and areas of test design the *Performance Descriptors* may help with.

Table 5-5: *ACTFL* **Performance Descriptor** ***key terms definitions and examples***

Term	Definition	Examples	May help with
Language domains			
Functions	Global tasks the learner can perform	Initiate, maintain, and end a conversation; Narrate and describe	Test task or test item type design
Contexts	Situations in which the learner can function	One's immediate environment	Test task or test item type design
Content	Topics which the learner can understand and discuss	General interest; Work-related	Test task or test item type design
Text types	That which the learner is able to understand and produce in order to perform the functions of the level	Words, phrases, sentences, questions, strings of sentences, connected sentences, paragraphs	Test task or test item type design
Language control	The level of control the learner has over certain features or strategies to produce or understand language		Design scoring criteria
Vocabulary	Vocabulary used to produce or understand language		Design scoring criteria

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Communication strategies	Strategies used to negotiate meaning, to understand text and messages, and to express oneself		Test task or test item type design, Design scoring criteria
Cultural awareness	Cultural products, practices, or perspectives the language learner may employ to communicate more successfully in the cultural setting		Test task or test item type design, Design scoring criteria
Modes of communication			
Interactional	Active negotiation of meaning among individuals	Speaking and listening (conversation); Reading and writing (text messages or via social media)	Test task or test item type design, Design scoring criteria
Interpretive	Interpretation of what the author, speaker, or producer wants the receiver of the message to understand	Reading (websites, stories, articles); Listening (speeches, messages, songs); Viewing (video clips) of authentic materials	Test task or test item type design, Design scoring criteria
Presentational	Creation of messages	Writing (messages, articles, reports); Speaking (telling a story, giving a speech, describing a poster); Visually representing (video or Power Point)	Test task or test item type design, Design scoring criteria

Note. Source for Term, Definition, and Examples columns: ACTFL *Performance Descriptors*, pp. 7-9

The remaining six pages of *Performance Descriptors* present tables which give two to three descriptors within three ranges (Novice, Intermediate, Advanced), according to the "language domains" of functions, contexts, etc., and still yet again cross-referenced with Interpersonal, Interpretive, and

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Presentational "modes of communication." See Table 5-6 for sample Novice level *Performance Descriptors.*

Table 5-6: ***Sample performance descriptors for speaking, listening, reading, and writing at a general Novice level for three "language domains"***

Performance Descriptor example (may help with stating a minimally passing score on a given scoring criterion such as "language control")	Language use situation (potentially speaks to test tasks or items)	Language domain (potentially speaks to scoring criteria)
Can usually comprehend highly practiced and basic messages when supported by visual or contextual clues, redundancy or restatement, when then the message contains familiar structures.	Interpersonal	Language control
Primarily relies on vocabulary to derive meaning from texts.	Interpretive	
Produces memorized language that is appropriate to the context; limited language control may require a sympathetic audience to be understood.	Presentational	
Able to understand and produce a number of high frequency words, highly practiced expressions, and formulaic questions.	Interpersonal	Vocabulary
Comprehends some, but not all of the time, highly predictable vocabulary, a limited number of words related to familiar topics, and formulaic expressions.	Interpretive	
Produces a number of high frequency words and formulaic expressions; able to use a limited variety of vocabulary on familiar topics.	Presentational	
May use some or all of the following strategies to maintain communication, able to: Imitate modeled words Use facial expressions and features Repeat words etc.	Interpersonal	Communication strategies

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

May use some of all of the following strategies to comprehend texts, able to: Skim and scan Rely on visual support and background knowledge etc.	Interpretive	
May use some or all of the following strategies to communicate, able to: Rely on practiced format Use facial expressions and gestures Resort to first language	Presentational	

Note. Source: ACTFL *Performance Descriptors*, pp. 14-19.

Three successive instruction-leading-to-test task examples for Interpersonal, Interpretative, and Presentational "modes" have been generously offered by Jean-Mari Hernandez Lopez, a working teacher/tester in Spanish at Westtown School in Pennsylvania (U.S.A.) (with permission, December 7, 2018). See Table 5-7.

Table 5-7: ***Three successive tasks reflecting Interpretive, Interactional, and Presentational "modes of communication"***

Interpretive Task - Tuesday, First Week, ninety minutes *You have heard of the fables of Jean de la Fontaine and you would like to read one of them. You go online to find one that you think is interesting and intriguing. You read it and you love it. The moral or message of the story speaks to you and you can't believe that it was written so long ago.*
Interpersonal Task - Tuesday, Second Week, ninety minutes *You decide to meet up with your friend to tell him/her about the story. To your surprise your friend already loves fables and he/she has a favorite. You strike a super interesting conversation concerning the fables that you enjoy.*
Presentational Task - Monday, Third Week, one hundred twenty minutes *After your unexpected conversation with your friend you feel eager to write your own fable. You go home (classroom!) and start writing a fable with three or four characters who act as humans and teach a lesson to the reader. You feel proud of your 250 to 300-word fable. At the end of you fable, write the lesson your story teaches.*

Depending on what a teacher/tester wanted to know, he or she could design scoring criteria for perhaps "vocabulary" (note Table 5-6 which strongly links vocabulary with Interpretive modes of communication). Table 5-8 shows possible scoring criteria for the Interpretive task in Table 5-7.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 5-8: ***Possible scoring criteria and passing criterion scores for Interpretive task based on "language domain" for "vocabulary"***

Comprehends highly predictable vocabulary from the website or the story:		
1	2	3
Not very often	Some of the time*	Often
*Criterion score for passing at Novice level		
Comprehends words related to familiar topics:		
1	2	3
Few	Limited*	Many
*Criterion score for passing at Novice level		
Recognizes and comprehends formulaic expressions:		
1	2	3
Few	Limited*	Many
*Criterion score for passing at Novice level		

The scoring criteria in Table 5-8 are suggested by the *Performance Descriptors* but there still remain many unanswered questions: What are "highly predictable vocabulary" items? What are "familiar topics" to these learners? What are "formulaic expressions" for learners? In other words, from what domain of language items would the teacher/tester draw his or her test content? Is there an accepted proficiency-based list of words, applicable to all unpredictable language use situations? Or, would the list come from course materials? How then would a course-based list correlate to a universe of unpredictable reading texts in the real world?

Finally, how would the teacher/tester measure learners' vocabulary within these domains, assuming they could be specified? Will they use fill-in-the blank test items? Matching items? Together, the ACTFL *Guidelines* (2012a) and *Performance Descriptors* (2012b) remain simply suggestive. It would take much extrapolation and thought on the part of teachers/testers to use the information at the level of practicality needed to design and use classroom tests.

Can-Do Statements. ACTFL has published a third freely available document, the *Can-Do Statements* (2018b). Multiple, redundant, downloadable documents make up the *Can-Do Statements*. Two are focused on here. One is an introduction which describes how the statements can be used and by whom. According to the ACTFL *Can-Do Introduction* (ACTFL & National Council of State Supervisors for Languages, 2018a) the *Can-Do Statements* are designed to be used by learners "to identify and set learning

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

goals and chart their progress," by teachers "to write communication learning targets for curriculum, unit and lesson plans," and by stakeholders "to clarify how well learners at different stages can communicate" (2018a, p. 1). A second document is the *Can-Do Statements* themselves (ACTFL & National Council of State Supervisors for Languages, 2018b: https://www.actfl.org/sites/default/files/CanDos/Can-Do_Benchmarks_Indicators.pdf).

The *Can-Do Introduction* offers test-relevant information on the now-familiar three modes of communication: Interpersonal, Interpretive, and Presentational. Interpersonal communication is where "learners interact and negotiate meaning in spoken, signed, or written conversations to share information, reactions, feelings, and opinions" (2018a, p. 1). Interpretive communication is where "learners understand, interpret, and analyze what is heard, read, or viewed on a variety of topics" (p. 1). Presentational communication is where "learners present information, concepts, and ideas to inform, explain, persuade, and narrate on a variety of topics using appropriate media and adapting to various audiences of listeners, readers, or viewers" (p. 1). The descriptions of interactional and presentational communication offer additional ideas to teachers/testers to create test tasks that learners must respond to. Of relevance to scoring criteria selection, the authors note that the *Can-Do Statements* are intended, among other things, to design "rubrics for performance-based grading." They are, however, not to be used as "an instrument for determining a letter or number grade" (2018a, p. 4). How teachers/testers are to reconcile these two contradictory statements is unclear.

The *Can-Do Statements* themselves appear as multiple pages of charts that are cross-referenced by proficiency levels and sub-levels (Novice-low, Novice-mid, Novice-high, etc.) and by Interpersonal, Interpretive, and Presentational modes of communication. A general *Can-Do* statement is offered, and then three more statements are given with more specific language use situations, such as "What can I understand, interpret or analyze in authentic information texts?" or "fictional texts" or "conversations and discussions" (2018b, p. 3). The following are for Novice-level learners engaging with Interactional communication language use situations. Elements from Table 5-5, which give language use examples for Interactional communication, have been added. See Table 5-9.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 5-9: Can-Do Statements *for Interactional communication at the three Novice sub-levels*

General statement for Novice-level learners: *I can communicate in spontaneous spoken, written, or signed conversations on both very familiar and everyday topics, using a variety of practiced or memorized words, phrases, simple sentences, and questions.*		
From the ACTFL *Guidelines, on Interactional communication* (2012a): Definition: Active negotiation of meaning among individuals Examples: Speaking and listening (conversation); Reading and writing (text messages or via social media)		
Answers the question: *How can I exchange information and ideas in conversations?*	Sub-level	Can-Do Statement
	Novice-low	*I can provide information by answering a few simple questions on very familiar topics, using practiced or memorized words and phrases, with the help of gestures or visuals.*
	Novice-mid	*I can request and provide information by asking and answering a few simple questions on very familiar and everyday topics, using a mixture of practiced or memorized words, phrases, and simple sentences.*
	Novice-high	*I can request and provide information by asking and answering practiced and some original questions on familiar and everyday topics, using simple sentences most of the time.*
Answers the question: *How can I meet my needs or address situations in conversations?*	Novice-low	*I can express basic needs, using practiced or memorized words and phrases, with the help of gestures or visuals.*
	Novice-mid	*I can express basic needs related to familiar and everyday activities, using a mixture of practiced or memorized words, phrases, and questions.*
	Novice-high	*I can interact with others to meet my basic needs related to routine everyday activities, using simple sentences and questions most of the time.*

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Answers the question: *How can I express, react to, and support preferences and opinions in conversations?*	Novice-low	*I can express basic preferences or feelings, using practiced or memorized words and phrases, with the help of gestures or visuals.*
	Novice-mid	*I can express my own preferences or feelings and react to those of others, using a mixture of practiced or memorized words, phrases, and questions.*
	Novice-high	*I can express, ask about, and react to preferences, feelings, or opinions on familiar topics, using simple sentences most of the time and asking questions to keep the conversation on topic.*

Note. Source: *Can-Do Statements* (ACTFL & National Council of State Supervisors for Languages, 2018b)

Taking the two *Can-Do Statement* documents described in this section, a teacher/tester might design a test task where learners give oral presentations on one of several concepts learned in class, such as the celebration of holidays in rural areas, or figuring out how to get from one city to another on public transport. Learners could then rate themselves on the overall question *How can I express, react to, and support preferences and opinions?* (Table 5-9, lower left). With some extrapolation, teachers/testers may ask Novice learners to focus particularly on corresponding ideas in the right hand column of Table 5-9: Are you using mostly memorized words and phrases? Are you using a mixture of memorized words, phrases, and questions? Are you using your own simple sentences most of the time?

Use of ACTFL materials by contributors to this book. Four contributors mention the ACTFL documents in their questionnaire responses. See Table 5-10.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 5-10: ***Contributors mentioning use of ACTFL materials***

Contributor	General purpose for using ACTFL materials	Used for task or test item format construction?	Used to construct scoring criteria?
Beatriz Garcia Glick "An Oral VoiceThread Test for First-semester French Language Learners in a U.S. University"	To test oral competence.	Yes.	No.
Annis Shaver "A Speaking Fluency Test for Intermediate-level German using a Rubric Based on Grice's Conversational Maxims"	The need to test in all four skills.	No.	No. The Ohio Department of Education offers oral test rubrics but she did not use them.
Borbala Gaspar and Margherita Berti "A Multiliteracies-oriented Project-based Assessment for Intermediate Foreign Language Italian Classes"	To address ACTFL's "Five C" and "World Readiness" standards which stipulate that learners connect to communities within and without the school.	Yes.	No.
Irina Drigalenko "A Written and Oral Russian Achievement Test for Beginning College-Level Learners"	To use as guidelines for test construction.	Yes.	No.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

It appears the ACTFL *Guidelines* were perhaps at best a general inspiration for the contributors' tests, particularly for testing speaking. It does not appear contributors used ACTFL materials to construct scoring criteria, which echoes earlier commentary on how difficult it is to extrapolate scoring criteria from freely available ACTFL documents. The ACTFL materials may appear as an element of the institutional context; teacher background; or current and past teacher education, classes, workshops components of the teacher theory model (Figure 2-1).

Conclusion. On the positive side, the free and downloadable ACTFL *Guidelines* (2012a), the *Performance Descriptors* (2012b), and the *Can-Do Statements* (2018b) together provide a somewhat internally coherent and suggestive guide for language instruction, and possibly testing. On the negative side, a lot of extrapolation and thought is required to use them, to the point that ACTFL has published books, for a cost, on how to interpret and integrate the ACTFL *Guidelines* into instruction and tests. Using the ACTFL *Guidelines* to design and use tests would be daunting if freely available materials from the ACTFL website were the sole resource. If teachers/testers had the time and money to take webinars, attend examiner certification workshops, and buy books, the ACTFL *Guidelines* may be more clearly a resource.

Common European Framework of Reference

The Council of Europe (COE), is a transnational governmental organization dating from the late 1940s with a political and cultural agenda to encourage "a common view of European Citizenship" (Fulcher, 2004, p. 253). Authorship of the two *CEFR* documents described here, as well as supporting research projects, are sponsored by the COE. *CEFR* is an expression of the language policy of COE, which is to encourage language learning in all the diverse languages in Europe as a means of personal mobility and mutual understanding (Council of Europe, 2001). In other words, if Europeans wish to cross borders to work or study or live elsewhere, they need interpersonal and professional skills in language and cultural exchange to succeed. To make language learning available to all, so the reasoning goes, one must be able state learning goals and set standards in ways that are mutually comprehensible between schools in different countries, and between educators and learners in different countries and language learning contexts (p. 3). Learning goals must be "worthwhile and realistic" (p. 3) and relevant to learners, and language learning methods should be employed that "will strengthen independence of thought...combined with social skills" (p. 4).

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

There are multiple, lengthy *CEFR* documents available online for free on the COE website, and two are focused on here. The first is the 262-page pdf document, the *Common European Framework of Reference for Languages: Learning, teaching, assessment* (Council of Europe, 2001). This will be referred to as 2001 *CEFR*. See the link: https://rm.coe.int/16802fc1bf The second is a 236-page pdf document, the *Common European Framework of Reference for Languages: Learning, teaching, assessment companion volume with new descriptors* (Council of Europe, 2018), which will be referred to as 2018 *CEFR*. See the link: https://rm.coe.int/CEFR-companion-volume-with-new-descriptors-2018/168078798 9

Both documents will be referenced in the remainder of this discussion. While they are related, they are not redundant. Further, teachers/testers contributing to this book (Gisela Mayr and Maria Martinez-Garcia) who cite being influenced by *CEFR* are likely to have been influenced by 2001 *CEFR*. 2018 *CEFR* has not been out long. As will be seen, 2001 *CEFR* approaches being a textbook, with an internally coherent syllabus. 2018 *CEFR* has less of this quality, instead providing hyperlinks to yet more documents. See Table 5-11 for the contents for both 2001 and 2018 *CEFR* with brief functional descriptions, particularly as they suggest resources for teachers/testers in terms of test task or test item ideas, scoring criteria, or determining how good a learner should be on a given scoring criterion (a standard).

Table 5-11: 2001 CEFR *and* 2018 CEFR *contents*

2001 *CEFR*		2018 *CEFR*	
Chapter/Section	Number of pages	Chapter/Section	Number of pages
1 The Common European Framework in its political context *Rationale for CEFR*	8 pages	Table of Contents	5 pages
2 Approach adopted *Identifies language learners as language users and social agents; Identifies assumptions about types of knowledge learners need to use language.*	11 pages	Preface *Identifies authorship; Gives list of schools in Africa; Asia; Central, North, and South America; and Europe that participated in validation of new descriptors on "mediation," "online interaction," and "plurilingual/pluricultural competence" among other things.*	10 pages

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

3 Common Reference Levels *Describes the six general levels of ability with multiple tables of descriptions for "understanding," "speaking," and "writing." Establishes scoring criteria for speaking such as "range," "accuracy," "fluency," "interaction," and "coherence."*	21 pages	Foreword and Introduction *Includes hyperlinks to additional documents such as "A Handbook for Curriculum Development and Teacher Education Concerning the Language Dimension in All Subjects"*	4 pages
4 Language use and the language user/learner *Elaborates on language use and identifies domains (personal, public, occupational, education) and situations of language use (locations, institutions, persons, objects, events, etc.), providing an analytic model rich in ideas for test task design or test content selection. There are descriptors for "overall oral production," "sustained monolog," etc., suggesting standards (How good does a learner have to be?)*	56 pages	Key aspects of the *CEFR* for teaching and learning *Reiterates and amplifies rationale for CEFR; Includes new "macro functions" of CEFR including Reception, Production, Interaction, and Mediation, each of which is cross-referenced with "Creative, interpersonal language use," "Transactional language use," and "Evaluative, problem-solving language use" (p. 31). These offer ideas on test task design or test content selection. Includes a summary of authors' model of competences. Reiterates and amplifies the idea of individual learner "profiles" where learners develop more in some areas than others along the six levels of the* CEFR *scale, depending on program aims and course objectives.*	19 pages

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

5 The user/learner's competences *Sets out* CEFR*'s theoretical model of the role of learners' knowledge in language use, including communicative competence. Offers descriptors by level for components of communicative competence, including grammatical, sociolinguistic, and pragmatic competence, suggesting scoring criteria.*	25 pages	The project to update and extend the *CEFR* illustrative descriptors *Technical account of how descriptors and scales for new areas (see "macrofunctions" above) were created and validated; Using* CEFR *for signed languages and with young learners.*	8 pages
6 Language learning and teaching *Establishes learners developing language use competences that are strong in some areas and weak in others, depending on the course objectives or program aims that have been formulated.*	25 pages	The *CEFR* illustrative descriptor scales *Describes the six general levels of ability with dozens of tables organized by macrofunction and language use situation. For example, under "production" are tables with descriptors for six levels for "sustained monologue: Describing experience" and "Written reports or essays." Under "Interaction" is "informal discussion" and "online conversation and discussions." Authors elaborate their model of "communicative language competences" and establish six levels of descriptors for thirteen different scales under the general components of linguistic, sociolinguistic, and pragmatic competence.*	108 pages

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

7 Tasks and their role in teaching *Models task performance in terms of task difficulty, conditions, constraints, etc. giving food for thought when considering test task or test item design.*	10 pages	Appendix 1 Salient features of spoken language at the *CEFR* levels *Includes elaborated descriptions of interactive speaking.*	3 pages
8 Linguistic diversification and the curriculum *Discusses policy decisions schools can make to encourage learning of multiple languages.*	8 pages	Appendix 2 Self-assessment grid (with online interaction and mediation) *Includes "I can" descriptor tables for all six levels for reception (listening and reading), interaction (spoken and written and online interaction), production (spoken and written), mediation (mediating a text, mediating concepts, mediating communication).*	4 pages
9 Assessment *Defines assessment broadly not just as tests but teachers' checklists, etc. Establishes validity, reliability, feasibility as three considerations for assessments. Confirms that assessments need to establish what learners can do and how well they can do it.*	19 pages	Appendix 3 Qualitative features of spoken language (expanded with phonology) *Two tables with descriptors for six levels by six criteria: Range, accuracy, fluency, interaction, coherence, phonology.*	2 pages
General bibliography	7 pages	Appendix 4 Written assessment grid *Descriptors of six levels by six criteria: Overall, range, coherence, accuracy, description, argument.*	2 pages

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Appendix A *Technical account of developing* CEFR *descriptors; Includes annotated bibliography.*	11 pages	Appendix 5 Development and validation of the extended illustrative descriptors *Describes a three year project to formulate and pilot the descriptor scales for "mediation" and "plurilingual/pluricultural competence"*	9 pages
Appendix B *Describes the research history of* CEFR.	8 pages	Appendix 6 Examples of use in different domains of online interaction and mediation activities *Multiple tables of descriptors by six levels (some with sub-levels) cross-referenced by language use situation, including "personal," "public," "occupational," and "educational."*	37 pages
Appendices C and D *Information on related scales and "Can-Do" statements.*	20 pages	Appendix 7 List of changes to specific 2001 descriptors	1 page
		Appendix 8 Sources for new descriptors *Bibliography with hyperlinks.*	4 pages
		Appendix 9 Supplementary descriptors *Additional tables on areas of sign language, phonology, and interpreting, among others.*	8 pages

Most striking is large number of tables of descriptors in 2018 *CEFR*. The single largest section is 108 pages long with dozens of illustrative descriptor scales. Many appendices (1, 2, 3, 4, 6, 9) are additional descriptors. The proliferation of tables may be due to an alignment of *CEFR* to the high-level theory of communicative competence, which has as a central feature consideration of language users' responses to different language use situations. Each table may represent some language use anticipated by the authors.

Communicative competence as a theoretical basis. Both *CEFR* documents, and more particularly 2018 *CEFR*, are well-referenced with extensive bibliographies. This suggests a concern on the part of the authors

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

to establish theoretical clarity for *CEFR*. It was argued earlier in this chapter that *CEFR* (2001 and 2018) takes as its basis the high-level theory of communicative competence. To reiterate Fulcher & Davidson (2007, p. 37), any model of language ability needs three dimensions: A model of knowledge ("what it means to know a language"), a model of performance ("underlying factors related to the ability to use language"), and actual language use ("how we understand specific instances of language use"). As mentioned earlier, both 2001 and 2018 *CEFR* embrace the concept of language use. 2018 *CEFR* compares, for example, "moving a wardrobe" with "tasks demanding greater sophistication of communication, such as agreeing on the preferred solution to an ethical problem, or holding a project meeting" (p. 29). Also present are models of knowledge ("General competences" and "Communicative language competences") and a basic model of performance ("Communicative language strategies")(Council of Europe, 2018, p. 30).

What does this mean in terms of how 2018 *CEFR* appears? As can be seen in Table 5-10, a full 108 pages are devoted to *CEFR* illustrative descriptor scales. In the "reception" section alone are 13 separate tables of descriptor scales: Five for listening comprehension ("Overall listening comprehension," "Understanding interaction between other speakers," "Listening as a member of a live audience," "Listening to announcements & instructions," "Listening to audio media & recordings")(Council of Europe, 2018, p. 54); Six for reading comprehension ("Overall reading comprehension," "Reading correspondence," "Reading for orientation," "Reading for information & argument," "Reading instructions," "Reading as a leisure activity")(p. 54); One for "Audio-visual (Watching TV, film & video)" (p. 54); and one for "Reception strategies" which is "Identifying cues & inferring" (p. 54). Each of the tables is an expression of an assumption that different language use situations make a difference to language users. They will use language according to their understanding of the language use situation demands. Each table, then, may represent a different set of resources for teachers/testers. See Table 5-12 for a comparison of two sets of descriptors, both for reading comprehension.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 5-12: ***2018* CEFR *descriptors for reading correspondence and reading as a leisure activity***

Level	Reading correspondence	Reading as a leisure activity
C2 Proficient user	Can understand specialised, formal correspondence on a complex topic.	Can read virtually all forms of the written language including classical or colloquial literary and non-literary writings in different genres, appreciating subtle distinctions of style and implicit as well as explicit meaning.
C1 Proficient user	Can understand any correspondence given the occasional use of a dictionary. Can understand implicit as well as explicit attitudes, emotions and opinions expressed in emails, discussion foruns, blogs, etc., provided that there are opportunities for re-reading and he/she has access to reference tools. Can understand slang, idiomatic expressions and jokes in private correspondence.	Can read and appreciate a variety of literary texts, provided that he/she can reread certain sections and that he/she can access reference tools if he/she wishes. Can read contemporary literary texts and non-fiction written in the standard form of the language with little difficulty and with appreciation of implicit meanings and ideas.
B2 Independent user	Can read correspondence relating to his/her field of interest and readily grasp the essential meaning. Can understand what is said in a personal email or posting even where some colloquial language is used.	Can read for pleasure with a large degree of independence, adapting style and speed of reading to different texts (e.g. magazines, more straightforward novels, history books, biographies, travelogues, guides, lyrics, poems), using appropriate reference sources selectively. Can read novels that have a strong, narrative plot and that are written in straightforward,

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

		unelaborated language, provided that he/she can take his/her time and use a dictionary.
B1 Independent user	B1 high Can understand formal correspondence on less familiar subjects well enough to redirect it to someone else. B1 low Can understand the description of events, feelings and wishes in personal letters well enough to correspond regularly with a pen friend. Can understand straightforward personal letters, emails or postings giving a relatively detailed account of events and experiences. Can understand standard formal correspondence and online postings in his/her area of professional interest.	B1 high Can read newspaper / magazine accounts of films, books, concerts etc. written for a wider audience and understand the main points. Can understand simple poems and song lyrics written in straightforward language and style. B1 low Can understand the description of places, events, explicitly expressed feelings and perspectives in narratives, guides and magazine articles that are written in high frequency, everyday language. Can understand a travel diary mainly describing the events of a journey and the experiences and discoveries the person made. Can follow the plot of stories, simple novels and comics with a clear linear storyline and high frequency everyday language, given regular use of a dictionary.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

A2 Basic user	A2 high Can understand a simple personal letter, email or post in which the person writing is talking about familiar subjects (such as friends or family) or asking questions on these subjects. Can understand basic types of standard routine letters and faxes (enquiries, orders, letters of confirmation etc.) on familiar topics. A2 low Can understand short simple personal letters. Can understand very simple formal emails and letters (e.g. confirmation of a booking or on-line purchase).	A2 high Can understand enough to read short, simple stories and comic strips involving familiar, concrete situations written in high frequency everyday language. Can understand the main points made in short magazine reports or guide entries that deal with concrete everyday topics (e.g. hobbies, sports, leisure activities, animals). A2 low Can understand short narratives and descriptions of someone's life that are written in simple words. Can understand what is happening in a photo story (e.g. in a lifestyle magazine) and form an impression of what the characters are like. Can understand much of the information provided in a short description of a person (e.g. a celebrity). Can understand the main point of a short article reporting an event that follows a predictable pattern (e.g. the Oscars), provided it is clearly written in simple language.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

A1 Basic user	Can understand short, simple messages on postcards. Can understand short, simple messages sent via social media or email (e.g. proposing what to do, when and where to meet).	Can understand short, illustrated narratives about everyday activities that are written in simple words. Can understand in outline short texts in illustrated stories, provided that the images help him/her to guess a lot of the content.
Pre-A1	Can understand from a letter, card or email the event to which he/she is being invited and the information given about day, time and location. Can recognise times and places in very simple notes and text messages from friends or colleagues, for example 'Back at 4 o'clock' or 'In the meeting room,' provided there are no abbreviations.	No descriptors available

Note. Source for Reading Correspondence descriptors, Council of Europe, 2018, p. 61. Source for Reading as a Leisure Activity descriptors, Council of Europe, 2018, p. 65.

Different descriptors, different language test tasks, and scoring criteria. Taking the example of learners of Japanese at an A2 level, it would not be difficult to extrapolate quite different-looking teaching activities and classroom test items or tasks inspired by the two different sets of descriptors. For reading correspondence, for instance, the authors state that the descriptors take into consideration the length and complexity of the message, how concrete the message is, and whether it follows a "routine format" (Council of Europe, 2018, p. 61). A "low" A2 level learner "can understand very simple formal emails and letters" (p. 61). This might result in a test task or classroom activity where learners use a Japanese-language website to order a small item pursuant to course goals to engage in written interactions. Learners then read a confirmation e-mail and offer a basic L1 translation (an example of "mediation," p. 112), or identify formal elements of the e-mail, such as the salutation or the specific language or constructions, suggesting that an order had been successful, or not. Teachers/testers might

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

write a scale to identify to what extent learners' L1 translation shows comprehension of the basic points of the L2 message, such as (Table 5-13).

Table 5-13: ***Scoring scale for translation as a comprehension measure***

1	2	3
L1 translation represents less than 20% of propositions in the L2 version. Level not adequate to show basic comprehension.	*L1 translation represents between 30 and 50% of propositions in the L2 version. Level adequate to show basic comprehension.*	*L1 translation represents over 50% of propositions in the L2 version. Level exceeds basic comprehension.*

For identification of key formal elements of an e-mail, the teacher/tester may ask learners to identify six elements, according to course objectives: Salutation, statement of topic of e-mail, statement of disposition of the online order, evidence of positive or negative language supporting disposition, statement of action needed, leave-taking. The teacher/tester may then decide that identifying at least four of the elements constitutes a passing score.

For reading as a leisure activity, descriptors are based on length of texts, "whether there are illustrations," text types ("simple descriptions of people and places" and "narratives"), and text topics ("hobbies, sports, leisure activities, animals") (Council of Europe, 2018, p. 65). Low A2 learners "can understand short narratives and descriptions of someone's life" and "can understand what is happening in a photo story…and form an impression of what the characters are like" (p. 65). Further, they "can understand the main point of a short article that follows a predictable pattern" (p. 65). Learners could be asked to select and re-read a short narrative story from among those they had been reading in class or for homework. They could then identify some basic points within the narrative, such as the language used to describe a character, or the elements of the story that provide the basic narrative. A teacher/tester might decide that learners need to identify and know the meaning of four out of six adjectives used in order to pass a test or quiz. A teacher/tester may ask learners to read a series of statements that are basic restatements of the main points of the narrative provided in a test text (either one that was read before, or one that is very similar to texts already read) and identify which are present in the story. A score of three out of five might be considered passing.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Use of *CEFR* materials by contributors to this book. Two contributors mention the *CEFR* in their questionnaire responses. See Table 5-14.

Table 5-14: ***Contributors mentioning use of* CEFR *materials***

Contributor	General purpose for using *CEFR* materials	Used for task or test item format construction?	Used to construct scoring criteria?
Gisela Mayr "A speaking skills test for high school learners of English in Southern Tyrol."	To adhere to institutional goals. To act as a point of reference to design test (task and text selection).	Yes.	Yes.
Maria Teresa Martinez-Garcia "A final project performance test for a Spanish conversation class at a Korean university"	To adhere to institutional goals. To decide cut scores.	Here the author cites the ACTFL *Guidelines* and an American-based workshop as the inspiration to write a "free production" task.	No.

While one contributor mentions *CEFR* as a significant resource for their test, the other does not. Yet for both contributors, *CEFR* seems to have an institutional presence, meaning that their schools, in two very different geographic locations, see *CEFR* levels as a signpost for learner achievement and program success. Gisela Mayr's contribution may be useful to readers who wish to know how she used *CEFR* as a resource for her test.

Conclusion. The two different tasks employ different text types (short e-mail versus a short narrative). While more prescriptions on texts are given in 2001 and 2018 *CEFR*, some of the same issues present in the discussion persist, as with the ACTFL documents. Mainly, from what domains would texts be chosen? Is there consensus on what rhetorical moves or grammatical constructions comprises more or less complex messages? Or what "routine formats" comprise e-mails of the type that A2 level learners can read? What specific resources would teachers/testers use, particularly those who have less experience selecting reading texts, to decide what

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

comprised "simple descriptions of people and places"? What would "simple" or more complex be?

Another issue is foreign language teachers who are accustomed to courses that specify grammar, vocabulary, phonology/graphology (called "the three elements" by *CEFR* authors, Council of Europe, 2018, p. 31) and/or the traditional "four skills" (p. 31), and how they might use the descriptors as resources for classroom testing. The descriptors for reading comprehension (Table 5-12), for instance, seem based on a model of reading comprehension that takes text genre, textual competence, and L1 textual knowledge into account. To many teachers, this author included, that is a version of heaven. For other teachers who are more accustomed to intensive reading practices and seeing L2 texts as a means of "delivering" L2 lexis and grammar, it is unclear what the descriptors offer. Might they ignore the reading comprehension descriptors in favor of descriptors for "grammatical accuracy" (Council of Europe, 2018, p. 133), or whatever descriptors have the most congruence with perceived course outcomes? This would come from influences of "Teacher background, or "Institutional context" in the Teacher Theory Model (Figure 2-1). One wonders if this may also be an effect created by having so many tables of descriptors from which to choose. One can focus exclusively on grammatical accuracy while still claiming they are using *CEFR* for teaching and testing.

To conclude, *CEFR* is in marked contrast to the ACTFL *Guidelines*, which are arguably based on the high-level theory of proficiency, and which assumes a proper target for learners in some undefinable and unknowable future use of language. Proficiency as a model does not take specific language use situations into account. Thus, the ACTFL *Guidelines* have few descriptors, giving an impression of greater coherence, but also fewer immediately apparent resources for teachers/testers for choosing test tasks or items, or formulating scoring criteria.

Target Language Use Analysis (TLU)

Target language use (TLU) analysis is an application of the high-level theory of communicative competence to the "problem" of writing tests that capture learners' communicative competence. TLU analysis is comprised of identifying a domain of language use tasks, selecting some of the tasks for consideration, and then using a checklist to describe the target language use situations. The target language use task analyses may then be used to compare real-life or instructional target language use tasks to proposed test tasks to establish authenticity. There are no contributors who specifically mention using TLU analysis to decide their test item formats or test tasks.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Nonetheless, contributors such as Myles Grogan, and Meredith Stephen and Meghan Kaiser explain in detail their thinking on how to match task topics and procedures to what learners have routinely experienced in class. It is interesting how one middle-level theory (TLU analysis) may be functionally instantiated in teachers' local theory, perhaps through an entirely different mechanism than formal courses teachers/testers have taken (see the model on teacher theory, Figure 2-1). The section on TLU analysis here is offered to readers as a resource from the formal testing literature.

Bachman and Palmer (1996), two language testing specialists, posited one of the more detailed models of communicative competence, and then proposed TLU analysis as a means of developing language tests "for eliciting instances of language use from which inferences can be made about an individual's language ability" (p. 45). In other words, if a test is written such that learners engage in language use very similar to that of "real-life domains" (what happens in the real world) or "language instruction domains" (what happens in classrooms), we have a better idea of what learners' performances on a test mean. If we want to know how well learners can communicate in, say, German, for the purposes of doing business, then we need write tests that capture what they need to be able to do. "In an office setting…language use tasks might include writing memos, preparing reports, answering and taking messages on the phone, and giving and following directions" (Bachman & Palmer, 1996, p. 45).

Of key importance is that communicative competence posits language use, and language use situations. Knowledge of language and ability to use language comprise only part of communicative competence. A third and inseparable component is what language users make of a language use situation: "because language use, by its very nature, is embedded in particular situations, each of which may vary in numerous ways, each instance of language use is virtually unique" (p. 44). TLU analysis assumes this as a working reality, but also assumes that "certain distinguishing characteristics of language use tasks" can be identified and used to write tests. The TLU analysis checklist is thought to capture these distinguishing characteristics.

A contrast needs to be made here between the high-level theories of communicative competence and proficiency that significantly affects test design. Proficiency assumes some future use of language that cannot be well defined. Proficiency is overall language ability regardless of curriculum (Center for Open Educational Resources on Language Learning, 2010). A teacher/tester might try writing a series of classroom tests that only might generally predict some level of ACTFL, because performance (what is done on classroom tests) is seen as different from proficiency (American Council

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

of Teachers of Foreign Languages, 2012b, p. 4). Communicative competence, however, and more specifically TLU analysis, assumes we can describe target language uses and write tests for them.

TLU example of a Korean-English bilingual language reading achievement test. Bachman & Palmer (1996) offer multiple example applications of their TLU analysis used to design real test tasks. One example, for first grade children, is distilled in Table 5-15. Bachman and Palmer (1996) first describe the domain of possible target language use tasks experienced by the children (the test takers) in their Korean-language classes. The learners are in a reading class using first grade level Korean language arts textbooks, and the domain of language use tasks includes learners decoding "individual morphemes and phonemes," identifying words in texts, reading aloud, matching words with definitions, answering teacher-fronted comprehension questions, and writing book reviews. The texts used are general interest topics and the genres presented include "prose, diaries, and letters" (p. 347). Characteristics common to all of the typically occurring instructional language use tasks were distilled using the TLU checklist (Table 5-15).

Table 5-15: ***Target language use analysis for a Korean-English bilingual classroom pursuant to designing a reading achievement test***

TLU categories	TLU task
Characteristics of the setting (where the target language use takes place)	
Physical setting	A classroom, and also learners' homes where they are commonly asked to do homework using their textbooks.
Participants	Learners are ages 5 - 7 and were born in the U.S. They are learning Korean so they may "retain and improve their native Korean language" (p. 345). The teachers are Korean and English speakers.
Time (duration) of task	Short

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Characteristics of input (what learners read or hear in the target language use situations)	
Format (overall description)	Spoken and written language taught in integrated fashion. There are discrete items (words and sounds) verbally highlighted from written texts but these "occur in a stream of naturally connected language use" (p. 346).
Channel	Visual (written texts)
Form	Receptive (written texts)
Language*	Grammatical and vocabulary features: Spoken and written Korean where "vocabulary and syntax are simple and adapted to the beginner level" (p. 347). Some teacher talk and peer talk is code switched between Korean and English. Textual and functional features: Functions are "widely varied" comprised of many subject areas such as language arts, science, ethics, etc. Topics of non-math and non- science written texts are "etiquette, school life, family, children's games…" (p. 347). Vocabulary and pictures in texts are "simple and restricted according to the child's age" (p. 347). Genre features: Prose, diaries, letters, poems, signs, notices Register features: "language between students and their elders is written in the honorific register, whereas language between peers occurs in the non-honorific register" (p. 347). Cultural reference features: There are references to Korean culture including holidays, food, customs (p. 347).
Types (of input)	Decoding individual sounds, "identification of written words," reading aloud, answering teacher's questions about the text, "simplified" book reviews with a "title, a theme, and topics," "matching written words to meanings" (p. 347).

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Characteristics of expected response (what learners say, write, or do in response to the target language use situations)	
Format (overall description)	Learners write short responses to comprehension questions in Korean, or code switched with English. Learners also respond non-verbally by nodding their heads, matching written items, and circling words.
Channel	Visual (writing, physical responses) and oral ("verbal expressions of a book's content" (p. 347).
Form	Productive (written and oral)
Language*	Grammatical, textual, vocabulary features: Spoken and written Korean and English, which is simple and "less diverse" than that of the input "which is typical for young learners." Functional features: Spoken and written Korean and English, which is simple and "less diverse" than that of the input "which is typical for young learners" (p. 347).
Types (of responses)	Short answers to questions, more extended production such as short book reviews.
Relationship between input and response	
Reactivity	Non-reciprocal (there is little interactivity between language users)
Scope	Learners are involved with texts, "negotiating the meaning of the graphic system" (p. 348)
Directness	Learners must draw on more than the input to respond, and bring to bear their "language ability, topical knowledge, affective schemata, and metacognitive strategies" (p. 348)

Note. *The section on "language" includes considerations of linguistic features including grammatical, textual, functional, vocabulary, genre, dialect, register, and cultural references (Bachman & Palmer, 2010, p. 296).

There are some TLU categories with meanings that are not immediately apparent. "Reactivity" is "extent to which the input or the response directly affects subsequent input and responses" (Bachman & Palmer, 2010, p. 79). The characteristic of reactivity may be reciprocal or non-reciprocal. In Table 5-12 on reading correspondence or leisure reading, the target language use tasks as they involve reading are non-reciprocal in that "there is neither feedback nor…interplay between language users" (Bachman & Palmer, 2010, p. 80). "Scope" is "the amount of input that must be processed in order

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

for the language user to respond" Bachman & Palmer (2010, p. 81). In Table 5-15 the young Korean learners are involved with texts, taking in a lot of input in order to complete tasks. "Directness" is "the degree to which the task can be…completed by referring primarily to the input" (Bachman & Palmer, 2010, p. 81). Again in Table 5-15, we see that learners have to draw on more than simply the input to be able to respond to the input.

Test designers used the target language use analysis to select two different tasks for two achievement tests: 1.) A monthly reading comprehension test with 8 - 10 items, and 2.) An end of course test comprised on a "simple book review" based on two chapters from learners' textbooks used in class (Bachman & Palmer, 1996, p. 349). Authors note that test takers may answer in either English or Korean, and with code switching (p. 350). There is little information on how the learners' performances are to be scored, which may not be an immediate purpose of TLU analysis. Such considerations may emerge from an examination of course goals or other sources. For instance, the authors identify the subskills for reading mentioned in a course syllabus for the children: 1.) "ability to recognize written vocabulary and comprehend its meaning," 2.) "ability to grasp the main idea of written words and sentences," and 3.) "ability to recognize specific details" (Bachman & Palmer, 1996, p. 349).

Conclusion. Target language use (TLU) analysis is an application of the high-level theory of communicative competence. It offers significant resources for deciding test tasks or test item types, relying on a principled procedure for identifying a domain of real-life or classroom language use tasks that learners experience, selecting some of the tasks, and then analyzing them. The analysis serves to help design test tasks that are authentic to what learners experience, and also to compare target language use tasks that learners experience and then what they experience while taking existing tests.

Multiple Literacies

Multiple literacies is an incipient middle-level theory and a movement in American foreign language education that posits goals for college-level foreign language learning beyond attaining communicative competence (Schechtman & Koser, 2008). Stated in practical terms, the movement aims to remove the division of beginning classes being devoted to speaking and learning the linguistic code of the second language, and intermediate and advanced classes being devoted to learning literature in the second language (Maxim, 2002, 2006; New London Group, 1996; Schechtman & Koser, 2008). Rather, second language learners should, from the beginning, use

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

authentic texts in the form of stories, poems, films, posters, and radio broadcasts to challenge learners "to consider alternative ways of seeing, feeling, and understanding things" (Pratt, Geisler, Kramsch, McGinnis, Patrikis, Rying, & Saussy, 2008, p. 290). Multiple texts become the focus of learning, and learners' guided engagement with texts comprises the main classroom activity in the course. Contrast this with many college-level foreign language classes in which the development of speaking skills take the main stage.

Stated in philosophical terms, multiple literacies suggest that the goal of foreign language education should be to help learners function in another language and thereby grow in their abilities to use "critical language awareness, interpretation and translation, historical and political consciousness, social sensibility, and aesthetic perception" (Pratt et al, 2008, p. 290). Asking learners to work with texts better reflects our current society, and prepares learners to learn to negotiate "the multiple linguistic and cultural differences in our society" (New London Group, 1996, p. 61). The New London Group, focusing on both first and foreign language education, note that young people must operate in workplaces vastly different than what came before, in which social relationships and teamwork are prized, requiring "informal, oral, interpersonal discourse" (p. 66). These new values have created "hybrid and interpersonally sensitive informal written forms, such as electronic mail" (p. 66). In sum, language cannot be seen as a single, unitary, standard form. Language has many forms, and is only one mode of representation of meaning.

Stated in theoretical terms, some applied linguists position multiple literacies as an extension of, or a rethinking of, the high-level theory of communicative competence. It is argued that one prevailing view of communicative competence "focuses primarily on linguistic achievements in the L2" (Schechtman & Koser, 2008, p. 309). This view refers to the "narrow" view of communicative competence mentioned earlier in the chapter. Rather, communicative competence should be broadened to include conceptions of learning where "language, culture, and literature are taught as a continuous whole" (p. 309). There should be stronger recognition of the role of language in creating and maintaining culture and "social viewpoints" (p. 309). This new, or enhanced, component of communicative competence might be called "translingual and transcultural competence" where learners learn about themselves through engagement with another language and culture (Pratt et al, 2008, p. 289). Foreign language learning should be done to enrich learners, not to relegate them to the status of being imperfect learners of a formal code (the L2). College-level learners come to the table with mature cognitive abilities, and knowledge of their own L1s. They can

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

handle the "language of texts" including features of "ambiguity," "metaphorical meaning," "overt, unstated objectives," and "open-ended situations" (Maxim, 2006, p. 20).

Testing and teaching in a multiple literacies framework. Various commentators offer teaching ideas and frameworks for getting learners to engage with texts, but offer fewer overt resources for testing. This is one reason the contribution by Gaspar and Berti for an Italian test is so fortunate. If a teacher/tester must shift from only testing learners' command of the formal linguistic code, what then do they test on? What test item formats, test tasks, and scoring criteria do they use?

There are some indirect responses to this question in the literature. Maxim articulates a four stage in-class procedure for guiding beginning learners' engagement with texts. The text he uses is an unsimplified authentic German romantic novel. Reading and texts are not relegated to the "if there is time" category of course goals. Learners read 1000 words of the text in pairs or small groups in each class meeting. In the first reading, the teacher guides learners to identify major events and pinpoint "the textual language used to convey these events" (Maxim, 2006, p. 22). In the second stage learners locate details in the text. In the third stage, learners write about the major events and details using the L2 forms in the text. Finally, learners are guided to discuss the "implications" of the text to understand its cultural and social meaning (p. 22). Kern (2008) offers a "poetry from prose" lesson plan with a similar four-step procedure for high beginning learners of French. First, learners choose and read a paragraph from an authentic prose text. Second, they read the paragraph and ask vocabulary questions. Third, learners write a poem from the words in the text. Finally, learners rewrite their poems and compare poems, "looking for different meanings that the words take on in different contexts" (p. 384). A teacher/tester would have to think through carefully what knowledge and skills learners develop while engaged in this carefully staged pair and group work, and then consider how to capture that knowledge and those skills in test tasks and scoring criteria.

Swaffar and Arens (2005) speak more directly to testing, noting that multiple literacies instruction requires changes in thinking about learner assessment. They suggest that teachers/testers working with multiple literacies recognize that learners do not look through texts for answers to questions about "isolated facts" which is what the authors think typical reading comprehension tests do. Rather, learners must identify "elements of content and context [in texts]" (p. 18). Learners' responses on tests would not be graded for "a teacher-imposed norm for correctness" (p. 18) but rather evidence that learners can express themselves with "situational

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

appropriateness" and with an apparent ability to "edit and self correct" (p. 19).

Gaspar and Berti's contribution. Gaspar and Berti call their test "A multiliteracies-oriented project-based assessment for intermediate foreign language Italian courses." Similar to Kern's staged activities described above, the test comprises multiple steps and stages done over a semester. There are five distinct stages with learners finding texts, then engaging with the texts and using the texts to create their own texts, in this case, a series of presentation slides and eventually a final presentation. The testing process also accords with Swaffar and Aren's (2005) value on learners' ability to be appropriate in the L2 and show an ability to self-correct and edit.

Each stage of the project has its own scoring system with scoring criteria, three levels, and descriptors for each level. Most remarkable are the scoring criteria for the various presentation slide drafts, which include: Introduction; Use of primary sources; Content depth and transitions; Content accuracy and comprehensibility; Originality and creativity; Pictures, videos and background choices; Clarity; Engaging resources and activities for peers; References slide; Length of notes to accompany each slide; and Knowledge gained and findings. One can see the value on appropriateness in several of the criteria (Engaging resources and activities for peers), and also the value placed on learners developing their critical thinking (Use of primary resources; Content depth and transitions). Learners' language is scored, just not using the traditional scoring criteria of "Vocabulary use," for example. Rather, vocabulary is scored as it serves learners' ability to present content. A middling learner on this criterion does the following: "The topic is sometimes covered throughout and other times covered superficially. Some of the vocabulary from the course is used. It seems that some information is lacking and tools are used only briefly."

Conclusion. To conclude, multiple literacies is an emerging middle-level theory with practical, philosophical, and theoretical dimensions. Seeing multiple literacies theoretically as an elaboration of communicative competence shifts focus from purely linguistic features of the L2, to learners' ability to engage with L2 texts with increasing insight on how language creates culture and social structures. Thus, learning the L2 is done for learner enrichment at a level commensurate with their already well-developed cognitive maturity and skills.

Chapter Summary

This chapter introduced high, middle, and low theories significant to teachers/testers making and using tests, including communicative

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

competence, and language proficiency. A High Middle Low Theory framework was introduced, which describes the breadth and functions of the many theories used in second language education. Second language use frameworks (middle theories) such as the *Common European Framework of Reference* and the ACTFL *Guidelines* were described as applications of high level theories communicative competence and proficiency. *CEFR* and the ACTFL *Guidelines* were explored and described in terms of the resources they may offer teachers/testers. The high and middle theories described in this chapter also partially formed the basis of the questionnaire that probed authors on their contributions to this book (Chapter Two). Authors' contributions were highlighted and related to the theories being offered in the chapter.

Further Reading

On Communicative Competence

Fulcher, G. (1998). Widdowson's model of communicative competence and the testing of reading: An exploratory study. *System, 26*, 281-302.

On Broad Conceptions of Communicative Competence in Tests

Venema, J. (2002). Developing classroom specific rating scales: Clarifying teacher assessment of oral communicative competence. *Shiken: JALT testing & evaluation SIG newsletter, 6*(1), 2-6.

On Narrow Conceptions of Communicative Competence in Tests

Barrette, C. (2004). An analysis of foreign language achievement test drafts. *Foreign Language Annals, 37*(1), 58-69.

Fox, C. (1993). Communicative competence and beliefs about language among graduate teaching assistants in French. *The Modern Language Journal, 77*(3), 313-324.

On Proficiency

Frain, T. (2009). *A comparative study of Korean university students before and after a criterion referenced test* (Unpublished master's thesis). University of Southern Queensland, Australia.

Harsch, C. (2014). General language proficiency revisited: Current and future issues. *Language Assessment Quarterly, 11,* 152-169.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

On ACTFL *Guidelines*

Kissling, E. & O'Donnell, M. (2015). Increasing language awareness and self-efficacy of FL students using self-assessment and the ACTFL proficiency guidelines. *Language Awareness, 24*(4), 283-302.

Liskin-Gasparro, J. (2003). The ACTFL Proficiency Guidelines and the oral proficiency interview: A brief history and analysis of their survival. *Foreign Language Annals, 36*(4), 483-490.

On the *Common European Framework of Reference*

Alderson, J.C. (2006). The CEFR and the need for more research. *The Modern Language Journal, 91*, 659-663.

Davidson, F. & Fulcher, G. (2007). The Common European Framework of Reference (CEFR) and the design of language tests: A matter of effect. *Language Teaching, 40,* 231-241.

Goodier, T. (2014). *Working with CEFR* can-do *statements: An investigation of UK English language teacher beliefs and published materials*. British Council ELT Master's Dissertation Awards Winner. Retrieved December 30, 2018 from: https://englishagenda.britishcouncil.org/sites/default/files/filefield_paths/working_with_CEFR_can-do_statements_v2_1.pdf

Runnels, J. (2014). An exploratory reliability and content analysis of the CEFR-Japan's A-level Can-do statements. *JALT Journal, 36*(1), 69-89.

On Target Language Use Situation Analysis

Bachman, L. & Palmer, A. (2010). Describing language use and language ability. In L. Bachman & A. Palmer, *Language assessment in practice* (pp. 33-58). Oxford: Oxford University Press.

Barrette, C. (2004). An analysis of foreign language achievement test drafts. *Foreign Language Annals, 37*(1), 58-69.

On Multiple Literacies

Swaffar, J. & Arens, K. (2005). *Remapping the foreign language curriculum.* New York: The Modern Language Association.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

CHAPTER SIX

PRACTICAL METHODS FOR VALIDATING AND IMPROVING TESTS

GRETA GORSUCH

What this Chapter is About

As can be seen in the tests and questionnaire responses in Chapter Four, contributors spent a lot of time and energy thinking through what they wanted to know about their learners, and then designing, refining, and improving their tests to fit their intentions. They took just as much care in thinking over how to score their tests, and to ensure they had a clear picture of what learners could and could not do. This chapter puts a name to this process, that of test validation. Test validation refers to the process by which a test writer creates a reasoned argument for the validity of his or her test (Davies, Brown, Elder, Hill, Lumley, & McNamara, 1999). In the context of this book, test validation refers to coherent actions taken by individuals, as teachers, to support the validity of their test and their scoring of the tests (Gorsuch & Griffee, 2018). Test validity has to do with whether a test measures what is intended, and whether scores from the test are being used for the appropriate decisions (Alderson, Clapham, & Wall, 1995; Bachman, 1990; Davies et al, 1999). In other words, how does a teacher/tester know what learners' test scores mean? If a learner gets a score of 20 on a 50-point criterion-referenced test, does it mean the learner knows 40% of the course content? It might seem a simple matter then to fail this learner (an achievement decision), or use the test score as evidence that some learners do not know the content (a diagnostic decision), or that multiple review sessions might be needed (a diagnostic decision). But then what is that content? Can it be clearly described? Is the content truly reflected in the test? And, does the test reflect only part of the course content? If so, which part?

In many language testing books and specialized journals this same basic area of interest is known as test validity. Test validity is an important and

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

robust research area (see for example Norris, 2008, and O'Sullivan & Weir, 2011). Test validity sometimes brings on images of professional testers hunched in front of their computers, doing statistical analyses beyond what ordinary teachers can do or want to do. Setting aside this image, ensuring that a test is valid, reliable, and fair is an important undertaking, and test validation is the means by which this is done. Test validation, in the context of this book, is a continuous action or a process which takes place throughout the life of a classroom test. Thus test validation strategies (methods used for test validation) are highlighted in this chapter. Many test validation strategies are surprisingly practical, and can be built into tests during the test writing, administration, scoring, and test score use stages (see Life of a Classroom Test Model, Figure 2-2, Chapter Two). As will be seen, many contributors to this book use test validation strategies at a variety of stages in the life of their tests.

This chapter will define three key concepts: Kunnan's test validity model including his four areas of concern, test validation strategies, and reliability. Examples for the three concepts will be given from contributors' commentaries in Chapter Four. These are offered so that readers may consider applications to their own classroom testing projects. Finally, there is a Chapter Summary and a section on Further Reading.

Definitions of Key Concepts

Kunnan's test validity model. Test validity models are an example of high-level theory (Table 2-1), alongside communicative competence. As such, test validity models are overarching, articulated and published, constantly and productively argued, and widely consequential in second language education (see Norris, 2008, in Further Reading for a proposed utilization-focused test validity model). Kunnan's 1998 model, based on the work of the late Samuel Messick (1989, 1995), reflected major shifts in thinking on test validity. In essence, test validity moved away from a narrow understanding, where a test writer would state what language knowledge or skill he or she intended to capture in a test and then present statistical evidence that the test in fact captured that knowledge or skill. Such evidence would include statistical measurement of a test's reliability, which is the extent to which a test captures some prespecified knowledge or skill consistently. Messick, as interpreted by Kunnan, kept this understanding intact, but added to it three more areas of concern, having to do with how tests, and scores from tests, impacted learners and other stakeholders (see Gorsuch & Griffee, 2018, in Further Reading, for a review of these shifts in thinking). This moved test validity, and thus test validation, from the near

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

exclusive domain of testing specialists to whoever writes and uses tests, including practitioners working in local contexts (Norris, 2008). See Figure 6-1 for the test validity model adapted from Kunnan (1998).

	Test interpretation	Test use
Evidence	A. Score interpretation	B. Test usefulness
Consequences	C. Stakeholder values	D. Social, and learning and teaching consequences

Figure 6-1. **A test validity model adapted from Kunnan (1998).**

"Test interpretation" and "Test use" (on the horizontal axis) interact with "Evidence" and "Consequences" (on the vertical axis) to create four quadrants: A. Score interpretation; B. Test usefulness; C. Stakeholder values; and D. Social, and learning and teaching consequences. See Kunnan (1998) in Further Reading for testing specialist point of view.

The four quadrants and test validation questions and focuses they suggest. Quadrant A, Score interpretation, is the idea that learners' test scores have to be interpreted and explained (validated). Test scores cannot be assumed to be objective or self-explanatory. In other words, teachers/testers need to be able to say what they think their test measures, design those ideas into a test, and develop evidence for it. Test validation questions or focuses arising from this quadrant might be: Does a test tell a teacher what they need to know? Does the test have accurate scoring procedures worked out? Has a test been scored the same way for all learners? Does a test have a relationship to a given curriculum? How much, or how well, is the test matched to the curriculum? These last questions refer to content validity, which is the degree to which a test resembles what learners experience in daily classroom activities. Content validity plays a large role in how contributors to this book believe they achieve test validity and test reliability.

Quadrant B, Test usefulness, has to do with "differences among learners, including experience with a test, academic background, test-taking strategies, second language background, age, and gender [which] may affect how well [learners] do on tests" (Gorsuch & Griffee, 2018, p. 232). Test validation questions arising from this area of concern are: Are there

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

identifiable groups of learners within a testing group? If so, should special testing conditions or scoring procedures be established to learner groups who may be unduly advantaged or disadvantaged by a test? See Norton & Stein (1998) in Further Reading for a plainly written account addressing test usefulness for learners in South Africa.

Quadrant C, Stakeholder values, has to do with learners' and teachers' values, and the values held by concerned others (stakeholders) such as parents, school administrators, future employers, and content specialists. This quadrant suggests test validation questions such as: Does a test have salience, or visibility, to learners? Do teachers and learners have conflicting values over a test or testing procedure? For instance learners may offer feedback to teachers that the writing test they took in class had them writing by hand, whereas they do most of their actual writing on a word processor. This raises the question: Could the writing test could be offered in a computer lab under supervised conditions? Teachers/testers may decide to make this change, only then to learn that the school does not allow students to take tests on computer due to previous concerns over cheating (see Dale Griffee's contribution). Finally Quadrant D, Social, and learning and teaching consequences, has to do with how learners', teachers', and other stakeholders' lives are affected by a test. This quadrant suggests many test validation focuses including: Does a test encourage students' learning? Does a test encourage both learning and teaching aligned with course outcomes? Does a test have institutional consequences for learners that teachers/testers may not intend or want?

Validation evidence for any of the quadrants (A, B, C, D) would be notes or reports of how the many questions given above, and other related questions, were considered and answered. Some of the questions suggest concerns that teachers/testers may not be able to practically address. For instance, it may not be possible to change a test in response to learners' or parents' values. But it is nonetheless important to take note of their values, or any other test validity issues, and to take them into account when interpreting the meaning of learners' test scores.

Examples of test validation evidence in contributions to this book. Many items in the Questionnaire (Chapter Two) probed contributors' insights on the validity of their tests. The questionnaire items were posed more from a standpoint of test validation work done as a process, and therefore, accomplished throughout the lifespan of a test (Life of a Test Model, Figure 2-2). In Chapter Two in Tables 2-2, 2-3, and 2-4 for example, the questionnaire items appear according to each stage in the life of a test, reflecting the belief of the author that test validity is arrived at in small pieces over time by practitioners. See the next section on test validation

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

strategies. Nevertheless, it is also possible to view questionnaire items, and contributors' responses to them, in the context of Kunnan's four areas of concern. See Tables 6-1, 6-2, 6-3, 6-4 below for examples of how contributors addressed test validity concerns in the context of Kunnan's four quadrants (1998).

Table 6-1: ***How contributors addressed test validity concerns in Quadrant A on score interpretation***

Contribution Kai-Ying Hsu, "A Chinese Achievement Test for Intermediate College-level Learners"	
Test validity concern	Responses
Does a test have a relationship to a given curriculum?	-Contributor referred to what learners did in class (lesson plans and handouts of the course) and matched them in the test. -Contributor identified "domains" of course content to sample from to write test items.
Contribution Annis Shaver, "A Speaking Fluency Test for Intermediate-level German Using a Rubric Based on Grice's Conversational Maxims"	
Test validity concern	Responses
Does a test tell a teacher what they need to know?	-Contributor designed scoring criteria that drew attention away from the "superficial influences" of pronunciation and grammar, and drew attention strongly to learners' ability "to participate in a conversation about a topic known to all concerned."
Contribution Maria Teresa Martinez-Garcia, "A Final Project Performance Test for a Spanish Conversation Class at a Korean University"	
Test validity concern	Responses
Does a test have accurate scoring procedures worked out?	-Contributor engaged in a two-step scoring procedure. She gave a score immediately after the performance test based on notes she took during the test. Then she randomized learners' videos and rescored them. -Contributor compared her first and second scores: "Normally, both evaluations were

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

	quite similar, although in some cases I realized I was being too harsh with some students."

Table 6-2: ***How contributors addressed test validity concerns in Quadrant B on test usefulness***

Contribution Sakae Onoda, "An English Collocation Knowledge Test for College-level Learners and Pre- and In-service Teachers"	
Test validity concern	Responses
Do tests and testing procedures take into account differences in how groups of learners respond to tests?	-Contributor created multiple versions of the test and administered them randomly to her three course sections. In this way, no one course section of students could be unduly advantaged taking the test. -Contributor kept the testing conditions the same for all three course sections to keep a level playing field.
Contribution Myles Grogan, "A Simple Speaking Test for an English-language University Communication Class"	
Test validity concern	Responses
Do tests and testing procedures take into account differences in how groups of learners respond to tests?	-Contributor identified two groups of learners who would unduly change the shared "interaction" score for other learners in their groups: "student who talk too much or too little." -Contributor taught learners the concept of "joint responsibility," which he also found was compatible with learners' own culture. -Class meetings leading up to the test were taken up in practice and topics very similar to the test.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 6-3: ***How contributors addressed test validity concerns in Quadrant C on stakeholder values***

Contribution Dale T. Griffee, "A Vocabulary Quiz for ESL Learners at an Intensive English Language School"	
Test validity concern	Responses
Do teachers and learners have conflicting values over a test or testing procedure?	-Students expected a multiple choice test item format but the instructor used a short answer format instead, to capture learners' ability to use a word in context where learners had to use the word in a sentence. The items were subjectively scored. This left open the door to learners arguing for higher grades, as the test was not objectively scored. Many teachers at the school made objectively scored tests to avoid conflicts with students.
Does a test have "buy in" from learners?	-Learners had told the instructor early in the course that they were weak on getting inferences from materials they read. Thus the instructor added a short response item to the test that captured learners' ability to infer meaning from a reading passage.
Contribution Borbala Gaspar and Margherita Berti, "A Multiliteracies-oriented Project-based Assessment for Intermediate Foreign Language Italian Classes"	
Test validity concern	Responses
Does a test have "buy in" from learners?	-Learners' projects arose out of their "personal interests" and through an iterative process of drafts and workshops, they became experts in a specific area of their choosing.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 6-4: ***How contributors addressed test validity concerns in Quadrant D on social, learning, and teaching consequences***

Contribution Irina Drigalenko, "A Written and Oral Russian Achievement Test for Beginning College-level Learners"	
Test validity concern	Responses
Does a test encourage students' learning?	-All marked unit tests were handed back to students to keep as a study guide for the final exam. -Instructor held a language learning lab session in advance of all unit tests for learners to practice. In addition, instructor had a review session before unit tests.
Contribution Gisela Mayr, "A Speaking Skills Test for High School Learners of English in Southern Tyrol"	
Test validity concern	Responses
Does a test have feedback potential for teaching and learning?	-Instructor used the test at the beginning of the school year, and involved her colleagues in administration and scoring. She thought the test might "raise others' awarenesses and change the prevailing way of handling and testing speaking competencies" at her school. -Instructor said the test scores helped her plan what to teach in following class meetings, such as including more activities that did not appear in the textbook such as "everyday speaking" and "current affairs."

Any of the responses in Tables 6-1, 6-2, 6-3, and 6-4, when noted and collected by teachers/testers during the design and use of a test, constitutes test validity evidence. Given evidence, then, Kai-Ying Hsu (Table 6-1) can rightfully use learners' test scores to decide whether learners learned from the materials and instruction the test was based on (achievement). This is what is meant by "interpretation" as it relates to test validity. Dale Griffee

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

(Table 6-3) can rightfully claim (and interpret) that learners' scores can be used to estimate whether they understand unit vocabulary well enough to use the vocabulary items in context, knowing full well learners will argue with him over the subjectively scored items. In fact, arguing over the subjectively scored items may further students' learning on this point.

Test validation strategies. Test validation strategies are small, coherent, and significant actions taken by teachers/testers to write tests to produce test scores and insights that create a reasonable basis for teachers' decision making (achievement and/or diagnosis). In essence, the strategies are efforts by teachers to understand what the test results are telling them. Test validation strategies are key to improving a test, and ensuring that scoring is accurate and fair. There are many examples of test validation strategies given above in Tables 6-1 through 6-4. In the tables, the examples are called "responses" to validation concerns. Nonetheless, as a high-level theory, Kunnan's test validity model (Figure 6-1) is abstract and may appear static. It may not be immediately clear to teachers/testers how the model can be used for the continuous actions or processes which typically take place throughout the life of a classroom test. Meredith Stephens and Meagan Kaiser, for example, revised their performance test's scoring criteria mid-semester to better reflect their course goals. Thus, in Tables 6-5 and 6-6 that follow, test validation strategies are offered for test planning and writing (beginnings of a test), test administration and scoring (the middle stages of a test). Validation strategies for the end stage of a test (using test scores) will be treated in Chapter Seven (Practical Methods for Setting Cut Scores). Tables 6-5 and 6-6 follow Tables 2-2 and 2-3 in Chapter Two which describes the framework used to create the questionnaire for contributors.

Many of the strategies in Tables 6-5 and 6-6 are found in the conventional testing literature (e.g., Brown, 2005; Gorsuch & Griffee, 2018; Hughes, 2010), but also find expression in the contributors' questionnaire responses. Thus, most strategies are illustrated by one example from contributors' comments, along with a direct quote from the contributors. The test validation strategies are also given a rationale stated in terms of what information a given strategy offers.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 6-5: *Test validation strategies at the test planning and writing stage*

Test validation strategy	Information offered	Contributor using the strategy
Overall test plan		
Describe the test. Consider: What are the general parameters of the test? How many items? How many subtests? What is the name of the course the test is for? Name the test.	-Defines time constraints. -Defines how much information can be collected. -Defines what kind of information can be collected (test scores, audio or video recordings, writing samples).	Kai-Ying Hsu: Four subtests with 18 total items Notable quote: "The main consideration was the time constraint. I planned that students will be able to finish the test in around 25 minutes to half an hour."
State the test purpose. Consider: What kind of decision is the test to be used for? How much course content coverage is intended? Is it a unit test or a quiz? A mid-term or a final exam?	-Gives insights on whether planned test is the appropriate length, format (paper and ink test, performance test, etc.).	Ferit Kiliçkaya: A final exam comprising 60% of learners' course grade (achievement) Notable quote: "Then, based on the contents of the coursebook and the corresponding weight of each structure in the book, I divided the number of questions by the number of the structures."
Deciding test content		
Link the test to course content. Consider: Does the test tap into specific course objectives? Does the test capture the same general balance or proportions of content that the learners experience in the course?	-Defines the extent to which a test has content validity. -Gives insights on whether an additional test or other form of learner data is needed to award a course grade, etc.	Myles Grogan: Final performance test matches topics and procedure of class meetings Notable quote: "Example textbook topics, like 'Human Migration'… become centred around a main speaking practice question, which is also used as the focus question for the test."

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Making test specifications		
Name the knowledge or skills a test is intended to capture. Consider: Talk through definition with a colleague, and/or show items; rethink and revise; consult literature.	-Gives insights on clarity of description. May help link ideas to high, middle, or low (teacher) theories.	Sakae Oneda: Test designed to capture learners' knowledge of verb collocations and multi-word units. Notable quote: "I checked with my colleague to confirm what I tried to measure…my colleague agreed with my suggestion by checking the test items."
Name the test item types (receptive/productive, fill in the blank, short answer, etc.) and match to what knowledge or skill is being measured. Write the items. Consider: Consult colleagues, revise.	-Gives insights on whether a test item type is appropriate to capture the knowledge or skills to be measured.	Yesica Amaya: Test designed to capture language knowledge and textual and functional knowledge. Notable quotes: "I prefer to use performance tests, matching, fill in the blank, and multiple choice depending on what I want to measure." "I think the [performance test subtest] test also captured textual and functional knowledge since students had to write a letter where they had to organize the information accurately."
Write test items		
Write more test items than needed. Consider: Show them to colleagues or learners (learners similar to those the test will be used on).	-Gives information on whether there are too few, enough, or too many items given what knowledge or skills need to be measured and given time constraints. -Helps teachers/testers winnow out or revise items that seem off-task.	Sakae Oneda: Wrote multiple forms of her collocation test so they could be administered quickly in multiple class meetings. Notable quote: "Since I decided to make more items than I needed and selected the best ones, I tried to write twice as many items as I needed to make multiple forms of the test."

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Decide and describe the test task (for performance tests). Consider: Consult curriculum, needs analysis, colleagues.	-Gives information on whether a test task is doable, appropriate, realistic.	Gisela Mayr: Created two performance tests, one monologic and one dialogic with learner pairs. Notable quote: "I wrote or prepared more test tasks than I actually used and chose the ones that seemed most suitable, after discussing them with my colleagues. The topics of the test tasks were chosen so as to be meaningful and interesting for the students."
Decide and describe scoring criteria (for performance tests). Consider: Consult literature, curriculum, colleagues.	-Gives insights on whether scoring criteria matches curriculum, purposes of test. -Gives insights on whether scoring criteria reflects theory, which will make the scoring criteria more easily describable, teachable, and usable for others (including learners).	Meredith Stephens and Meagan Kaiser: Revised scoring criteria on an oral summary test after a mid-term, based on course goals and communicative competence theory. Notable quote: "Also, setting up the rubric in a way that more closely mirrored the arc of the story seemed like a clearer and more effective way to explain to students what we hoped for them to accomplish."
Identify test taker bias and values. Consider: Consult learners. Compare test content and procedures to that which learners experience in lessons.	-Gives information on potential sources of test-taker bias and differing values, and create responses to them.	

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Plan scoring		
Plan scoring. Consider: Keep notes, consult colleagues.	-Gives insights as to how objective items will be scored (partial credit or no?), and whether the total scores will be easy to interpret. -Also gives insights on how many scoring criteria can be used for a given performance test task, balanced with time constraints.	Irina Drigalenko: Limited scoring criteria to three for an essay/dialog performance subtest, decided the reasons why points would be deducted from an assumption of 25 points possible. Notable quote: "For example if the essay was too short, additional points were taken off." "…each roughly written mistake was penalized by a whole point, less-rough mistakes and typos cost the student a half-point."

The fact that just one contributor example is given for most strategies above is not meant to suggest that some strategies are done more, or less, than other strategies. This book is not intended to characterize what teachers/testers do, nor does it have the aim of creating claims generalizable to all practitioners. Nonetheless, some validation strategies were mentioned by every contributor. Among them was carefully planning the number of items or tasks learners were asked to do. This was driven by a nearly universal concern for how long the test would take to administer. Most contributors administered their tests during regular classroom meetings and were thus constrained by time limitations. Another validation strategy mentioned by virtually every contributor was matching test items, test tasks, or scoring criteria to classroom topics and content, and course objectives. This seems to refute Barrette's (2004) findings which suggested teachers in a Spanish program did not use items or subtests that reflected what learners did in class.

See Table 6-6 for test validation strategies done at the test administering and scoring stage.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 6-6: ***Test validation strategies at the test administering or scoring stage***

Test validation strategy	Information offered	Contributor using the strategy
Design and assemble the test		
Check test layout on the printed page. Consider: Do all learners have an equal chance to do well? Keep notes on insights, any changes made during this stage.	-Gives insights on whether layout of the test is confusing for some or all learners; whether the artwork is clear.	
Decide whether to make two or more versions of the test. Consider: Consult colleagues, learners similar to learners taking the test.	-Gives insights as to whether different test versions are equivalent in difficulty, or sampling the same content.	Sakae Oneda: Created multiple forms of her short test. Notable quote: "I created multiple versions of the test by piloting it with groups of learners who demonstrated similar learner characteristics as the target students in terms of nationality, proficiency, and age."
Peer review the assembled test. Consider: Show test to colleagues, learners similar to learners taking the test. Keep notes.	-Shows whether test directions are clear, whether there are critical typographical errors.	Kai-Ying Hsu asked her testing instructor to peer review her achievement test before administration. Notable quote: "Major changes were made…the test instructions were modified to be more explicit."
Check audio files, video files, texts that learners respond to for the test. Consider: Keep notes, make corrections where needed.	-Gives information on whether audio/video files/texts are what are needed and appropriate for the test.	Kai-Ying Hsu removed background static from the video clip learners were to respond to during the test. Notable quote: "…the listening audio file [track] was edited to reduce distracting background noise."

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

For computer-mediated tests, check that technology is in good order. Plan work arounds if one or more computers malfunction. Consider: Do learners have adequate practice using the technology? Keep notes, schedule practice sessions if needed.	-Gives insight as to whether some learners' test performances may be thrown off by non-functional technology. -Gives advance notice if learners have not had sufficient practice using the technology to mediate or record their responses.	Beatriz García Glick: Learners recorded their performances on their own computers using a university-wide software platform. Notable quote: "I met with my class in a Technology Laboratory and we reviewed the goals of the test, the ways of uploading the information, and began a practice test in class to show them how to upload the information."
Reproduce the test		
Check that test photocopies, performance test texts or directions photocopies, rater sheet photocopies, etc. are in order and legible. Consider: Keep notes plus a clean copy of the test.	-Ensures that all learners taking the test can do so without undue difficulty or distraction.	
Make plans for security of test photocopies, performance test texts. Consider: Note that tests were kept in a secure location.	-Addresses one source of learner bias, that some learners may know the test content ahead of time.	Juliana Jandre and Vander Viana: Were concerned about test security but were more focused on security during the test, especially for the reading comprehension items which could be answered with stock answers. Notable quote: "…students were asked to skip a row from one another when choosing their seats…the test was invigilated by a school teacher, and students were asked to leave once they had finished the test…"

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Administer the test/Pilot the test		
Keep notes on how learners/testers/scorers interact with the test in real time conditions. Consider: Collect learners' comments. Keep notes on changes to test or testing procedures or rating procedures made due to insights on the administration/pilot.	-Give insights on whether test and testing procedure are workable; whether the test items work; whether the test task and scoring criteria can be accomplished/used in real time.	Ferit Kiliçkaya: He had two M.A. level learners assigned to him as student teachers take the test. Notable quotes: "I obtained my MA students' views on the complexity and difficulty of the test." "Based on their responses, I introduced several changes such as simplifying the language and the selection of the examples."
Rater training		
Conduct rater training. Consider: Keep notes on training procedure, questions or issues that came up. For instance scoring criteria that are specific and with a stronger basis in theory are easier to teach to other raters/colleagues. Ask other raters to evaluate the training.	-Gives insights to clarify scoring criteria, procedure.	Myles Grogan: Due to constraints he was unable to ask colleagues to help rate the test. Notable quote: "Getting colleagues to score tests in the context may not be impossible but it is far from practical. I have, however, had other teachers check rubrics, or discussed aspects of how [the] assessment may be made fairer."
Plan for intra-rater reliability procedure for performance tests. Consider: If working alone, rate learners' performances once, then set aside for one week and re-rate in randomized order without reference to first score.	-Gives insights on consistency of rating over time. -Provides potential insights on clarity of scoring criteria.	Maria Teresa Martinez-Garcia: Used a modified intra-rater scoring approach where she scored learners during their performances and then re-scored them at home based on notes she took during the performances. Notable quote: "Once I was back at home, I evaluated everybody's performance a second

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

		time (the second occasion), this time randomizing the order in which I reviewed and presentations/videos and I looked at my own notes."
Score test		
Downing (2006) suggests a two-step scoring procedure. Consider: Score once, create a test key based on the first scoring. Score a second time using the test key. Consider alternate answers which may not be expected but are still appropriate. Consult colleagues on alternate answers.	-Provides information on consistency and fairness of scoring.	Juliana Jandre and Vander Viana: Their test was scored by a teacher and pre-service student teachers. When it was discovered that the student teachers were marking too strictly, the tests were re-scored by the teacher. Notable quote: "When correcting Item 4-a for instance, they [the student teachers] considered it wrong when students liked the advertising campaign which mocked the Brazilian soccer coach, disregarding their [the students'] argumentation."
Blind the scoring. Consider: Cover learners' names as you score.	-Provides evidence of consistency and fairness of scoring.	
Change order of scoring. Consider: Shuffle learners' tests from order in which they handed in the tests.	-Provides evidence of consistency and fairness of scoring.	Taichi Yamashita: Scored tests both in order of submission but also backwards.
Put scores into a spreadsheet for specialized item analysis: Item facility, B-index, Difference index. Consider: Consult Brown (2005); Gorsuch &	-Gives insights on how well individual items on an objectively scored test functioned.	Taichi Yamashita: Calculated item facility and B-index on each of the objectively scored test items. Notable quote: "This process helped me identify potentially

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Griffee (2018) for specific instructions.		problematic test items, whose B-index is negative, and a potential underlying reason. For instance, I found that sometimes the discourse texts for the listening comprehension may have been too long for students to follow easily."
Estimate reliability		
Calculate Cronbach's alpha for objectively scored tests. Consider: Consult Brown (2005); Gorsuch & Griffee (2018) for specific instructions and formulae.	-Provides evidence of internal consistency and thus the theoretical clarity of the test.	Dale Griffee: Calculated Cronbach's alpha and KR-21 reliability coefficients. Notable quote: "I interpret the reliability coefficients as being minimally adequate for a classroom test."
Calculate Pearson Product Moment between raters for a performance test. Consider: Consult Brown (2005); Gorsuch & Griffee (2018) for specific instructions and formula.	-Provides evidence of consistency and fairness of scoring between raters, or between rater occasions (intra-rater scoring).	

Teachers/testers are not often called upon to write formal test validation reports on their tests, but as contributors' comments from Chapter Four and Tables 6-5 and 6-6 suggest, they nonetheless use test validation strategies at different points in the life of a test.

Some of the validation strategies were not apparently used by contributors, such as calculating inter-rater reliability or checking a printed test's layout. The former may echo the theme of teachers working alone, or may be the result of some contributors writing objectively scored criterion-referenced tests (not performance tests). Teachers/testers may also think of reliability differently than a technical, statistical understanding of the topic. The latter (checking a printed test's layout) may represent a process beneath conscious thought, and seen as just another straightforward photocopying job done under time pressure. But see Brown's (2005) specific suggestion on ensuring that everything a learner needs to answer a test item is all on

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

one page. A quick check of page layout would prevent half the answer choices being on a separate page for a multiple choice item or a matching item subtest. The concern here would be that learners would have to flip back and forth between pages to answer, or that some learners would not recognize that there are more answer choices than they thought. Either might call into doubt whether learners' test scores reflected their true abilities. Brown's suggestion on test page layout suggests the existence an easy-to-do yet perhaps under-represented test validation strategy.

Several themes emerged which were not specifically probed for by the questionnaire presented in Chapter Two. One theme is that of "revision in progress" where contributors reported changing scoring criteria mid-term, or changing test administration procedures due to shifting, contingent ideas on what learner knowledge or skills they wanted to tap into. This suggests that teacher theory (Figure 2-1) results in actions that pertain to test validity based on teachers'/testers' background, current and past education, the institutional context, and problems that needed to be solved. Another theme was how many contributors had to work alone. Some contributors were the only speakers of the language they were teaching at their institutions. Other contributors felt their colleagues were too busy to help, or that their colleagues would introduce unwanted and unconsciously applied scoring criteria into the scoring process.

Counterbalancing this was commentary suggesting that a culture of testing mentorship was alive and well, and supportive of test validity in the second language education field. Some contributors reported being assigned pre-service student teachers to mentor, who were then actively engaged in taking pilot tests and scoring tests. Some student teachers were guided and corrected in the process. At the very least, the data the student teachers offered was apparently seen as useful by the contributors, a stance which may be key to a productive mentorship. Other contributors apparently worked with instructors junior to them, or colleagues who were newer to testing, and together they created and administered tests. Still other contributors had enrolled in a testing course as graduate students, and worked directly with their course instructor on their tests.

Reliability. Test reliability may be broadly seen as to whether a learner's score on a test can be trusted. In other words, would all learners with the same level of knowledge or skill in a class get the same basic score? Alternatively, does everyone in the class get the same opportunity to show their ability, whether they have high or middle or low ability? Consider: Are the test items clearly stated? Are the directions clear to everyone? Can everyone hear the audio on a listening test equally well? Does everyone have equal times in which to speak, or write? Are the same scoring criteria

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

applied to everyone in the same way? Has the teacher added up the scores accurately? Has everyone had the same amount of training on the technology used to take a test? Ultimately, some learners will do well or poorly on a test. Reliability poses the question of whether learners' scores can be trusted to transparently indicate how well or poorly learners did. Learners should not get a low score because test directions were confusing, or because the teacher unknowingly scored harder on the grammar criterion (where learners had trouble) but then ignored the vocabulary criterion (where learners did well). In sum, a learner's score on a test should be the result of what they know and can do in the second language in terms of the course content, but not the result of irregularities or design faults in a test, testing procedure, or scoring procedure. It can be argued that reliability is a cornerstone of test validity, and is well represented in Quadrant A of Kunnan's (1998) four areas of test validity concerns (Figure 6-1).

There are many sources on reliability explored in technical and statistical terms. Any general second language testing book will have informative chapters on reliability (see for example Brown, 2005; Gorsuch & Griffee, 2018). And in fact, many contributors to this book evidenced this technical knowledge of reliability, but at the same time relied on other means of arguing for the reliability of their tests. Their commentary suggested an overarching concern for reliability and a wide array of methods for attaining consistency and trustworthiness of their test scores. See Table 6-7 for a sample.

Table 6-7: ***How contributors addressed test reliability***

Contributor	Test type	Approaches to reliability
Myles Grogan, "A Simple Speaking Test for an English-Language University Communication Class"	Subjectively scored performance test	-Did not try to achieve "psychometric reliability," rather focused on "fairness." -Ensured test topics, performance test task, scoring criteria were consistent throughout the semester with daily classroom learner activity (content validity).
Juliana Jandre and Vander Viana, "An English as a Foreign Language Test for Reading, Writing, and Cultural Diversity	Subjectively scored criterion-referenced test	-Sample size was too low to do statistical reliability estimates. -Asked student teachers to check the test for clarity. -Ensured all students had similar testing conditions. -Answered students' questions during test period.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Awareness for High School Students"		
Gisela Mayr, "A Speaking Skills Test for High School Learners of English in Southern Tyrol"	Subjectively scored performance test	-Statistical reliability hard to achieve with complex performance tests. -Ensured scoring criteria were clear. -Hired suitable raters/instructors who were proficient in the language being tested, and who did not know the learners. -Ensured consistency by using same raters for all learners in a testing session. -Scheduled the same amount of testing time for all learners. -Some learners did not have the same texts to prepare from, and that concerned her, but the tests were not graded and were used for course content planning.
Annis Shaver, "A Speaking Fluency Test for Intermediate-level German Using a Rubric Based on Grice's Conversational Maxims"	Subjectively scored performance test	-Ensured she used the same scoring rubric throughout the semester (learners did the performance test multiple times). -She reviewed learners' scores longitudinally and found overall improvement over the semester but no wide variations within the semester on specific test occasions. This suggested to her she was applying the scoring criteria consistently.
Yesica Amaya, "La Historia de la Pola: An Achievement Test for Original Content-based Materials for Beginning Learners of Spanish"	Objectively and subjectively scored criterion-referenced test	-She had a small sample size with few items. -Did a pilot test with masters and non-masters group. -Did item analysis with a focus on Difference Index (DI), comparing learners' answers on items before and after the treatment. -Did a peer-review of the test before administration.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

From the responses it was clear contributors were concerned about the soundness of the scores from their tests, and that they used many methods for achieving, for them, a reasonable sense that their tests were reliable. Many of these methods are mentioned in general testing books, and are characterized as being part of good test design and planning practices, including piloting and then revising a test, and writing clear scoring criteria.

Chapter Summary

Test validation is the all-important process of supporting test validity, or the trustworthiness of a test and the scores produced by a test. Test validation strategies underpin teachers'/testers' understandings of what learners' test scores mean, and thus what kinds of decisions teachers/testers may be comfortable making on the basis of the test scores. The term test validation strategy was defined, and examples were given from contributors' commentaries found in Chapter Four. Many of the strategies are applicable to every day testing projects that teachers/testers engage in. Test validation was also explored in terms of Kunnan's test validity model, a high-level theory more associated with testing specialists and formal research in testing. Yet it was found that contributors, as testing practitioners, had action-oriented responses to issues raised in Kunnan's (1998) four quadrants of concern. The chapter also explored test reliability, which has to do with the accuracy and consistency of the scores a test can produce.

Contributors had an overarching concern over reliability, and knew that it could be statistically estimated, with one contributor going so far as to call it "psychometric reliability." Yet many contributors turned to other methods for achieving reliability, including good scoring criteria design, careful rater selection, ensuring all learners had the same test conditions, and establishing content validity of their tests.

Further Reading

The following are about test validity and test validation, as explored by language testing specialists and language testing practitioners.

Content Validity

Siddiek, A.G. (2010). The impact of test content validity on language teaching and learning. *Asian Social Science, 6*(12), 133-143. Retrieved May 25, 2019 from: https://files.eric.ed.gov/fulltext/ED574721.pdf

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

General Testing Books or Resources with Sections on Test Validity and Reliability

Alderson, J.C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation.* Cambridge: Cambridge University Press.

Brown, J.D. (2005). *Testing in language programs.* New York: McGraw Hill.

Center for Open Educational Resources on Language Teaching (2010). *Lesson 4: Key ideas in assessment*. University of Texas at Austin. Retrieved May 25, 2019 from: https://coerll.utexas.edu/methods/modules/assessment/04/

Gorsuch, G. & Griffee, D.T. (2018). *Second language testing for student evaluation and classroom research.* Charlotte, NC: Information Age Publishing.

Hughes, A. (2010). *Testing for language teachers* (2nd Ed.). Cambridge: Cambridge University Press.

Test Usefulness (Quadrant B of Kunnan's Model—how tests interact with different groups of learners)

Norton, B. & Stein, P. (1998). Why the "Monkey's Passage" bombed: Tests, Genres, and Teaching. In A.J. Kunnan (Ed.). *Validation in language assessment* (pp. 231-249). Mahwah, NJ: Lawrence Erlbaum Associates.

Test Validity Models

Kunman, A.J. (1998). Approaches to validation in language assessment. In A.J. Kunman (Ed.). *Validation in language assessment* (pp. 1-16). Mahwah, NJ: Lawrence Erlbaum Associates.

Norris, J.M. (2008). *Validity evaluation in language assessment.* Frankfurt, Germany: Peter Lang.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

CHAPTER SEVEN

PRACTICAL METHODS FOR SETTING CUT SCORES AND MAKING DECISIONS

GRETA GORSUCH

What this Chapter is About

This chapter is on practical methods for setting cut scores. Many teachers/testers score their own tests, but as seen in Chapter Three on Criterion-referenced Tests and Performance Tests, they must also make decisions with tests (diagnosis or achievement). And for that, they need to set cut scores. For diagnosis: At what point, in terms of test scores, does the teacher/tester decide that learners do not know the content and that additional work and instruction is needed? For achievement: If a learner gets a particular score on a test, what does that mean, exactly? What does a score of 80% suggest about what a learner knows and can do? Then there are questions of placing learners' test scores into categories or "grades": At what point does a learner pass or fail a test? How many points on a test determine what grade (A, B, C, etc., or some other non-letter grade system) a learner gets on a test? What makes a score of 80% a C, or a B, or some other grade? Finally, there are questions as to what is done, institutionally, with the grades: What is the relationship between the test grade and the course grade--is a test consequential to the course grade, or not?

This last question points strongly to the teacher theory model in Chapter Two (Figure 2-1), in particular the role the institution plays in teachers'/testers' cut score and grading decisions. Contributors to this book work in widely different institutions, and as will be seen, the institution sometimes plays an overt role through requirements that just a few learners get As, but more learners get Bs, and some number of learners in between get Cs. This may be the case even when learners as a group get similar scores on a test. This is evidence of norm-referenced test thinking (see Chapter Three), which creates an uneasy fit for the criterion-referenced tests and

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

performance tests that make up the contributions to this book. Other contributors offer evidence of less clearly stated requirements pointing to a milieu or a culture of grading formed by their institutions, their personal teaching background, or past workshops or coursework they have taken.

Contributors' comments on making cut scores can be set against a larger educational backdrop of significant movements in the past three decades toward standard setting, performance descriptors, and criterion-referenced tests (Hambleton & Sireci, 1997; Livingston & Zieky, 1982; Shepard, 2000a; Zieky & Perie, 2004). These are positive changes in that learners are compared to course content or to course-based descriptors of language knowledge and skills developed by educators (criterion-referenced test thinking), as opposed to being compared to each other (norm-referenced test thinking). This chapter is offered to help teachers/testers navigate these movements. Authors working in these newer traditions report on cut score methods that are surprisingly practical, useful, and which facilitate fairness and clarity (important elements of test validity). Another name for cut score methods coming out of this newer tradition is standard setting.

Practical methods of setting cut scores and standard setting are given and adapted to both pass/fail decisions and grade decisions. They offer ways for teachers/testers to move beyond intuition. Key definitions are given, and where appropriate, examples are given from contributors' tests and test commentaries. The concepts defined are: Tests given for formative purposes, tests given for summative purposes, cut score, and standard setting. Finally, there is a Chapter Summary and a Further Reading section.

Definitions of Key Concepts

The first two terms defined here have to do with one issue mentioned earlier, that of whether or not test scores are reported to an institution. It is argued here that tests can be given for non-institutional purposes (formative) and also for institutional purposes (summative). Yet whatever the purpose, teachers/testers still need to set cut scores in order to make decisions with the scores.

Tests given for formative purposes. Typically the term "formative" is used in evaluation, which is a specialized, applied research field which seeks to judge the worth of a second language course or program (Griffee & Gorsuch, 2016). The term formative can also be applied to tests given to learners, and is done "during the course for the purpose of improving, illuminating, and clarifying" (Gorsuch & Griffee, 2018, p. 322) how well a course, or in this case a learner, is proceeding. Tests given for formative purposes may be tests used for diagnostic decisions, where the teacher/tester

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

uses learners' scores as a means to plan review or additional instruction. Perhaps learners did not do well on particular test items, subtests, or a test task. Teachers/testers may also give tests or quizzes during a semester or term to motivate or focus learners, without specific plans to review or re-teach content. And finally, teachers/testers may use achievement tests for formative purposes, to see how much students have learned, again, without specific plans to review or re-teach content. The point is that tests given purely for formative purposes are not reported to the institution for the purpose of awarding a grade. Tests given for the purpose of awarding a grade would be tests given for summative purposes (see the next defined term). As will be seen in this chapter, and also in Chapter Eight (Practical Methods for Using Tests for Teaching and Learning), many contributors made and used their tests for both formative and summative purposes. However, four contributions fall into the category of tests given for purely formative purposes, where the test was scored but not reported to a school. See Table 7-1.

Table 7-1: ***Tests by contributors given for formative purposes***

Contribution	Purpose	Cut score used and standard setting method used to set cut score
Kai-Ying Hsu, "A Chinese Achievement Test for Intermediate College-level Learners"	"[the scores] were mainly used for my research purpose to investigate whether authentic materials are effective to promote language learning." "It was a non-graded test thus it did not affect learners' grades."	The contributor wanted to set a cut score using the contrasting groups method to know if learners had learned from her authentic materials but felt she could not with so few students (five).
Gisela Mayr, "A Speaking Skills Test for High School Learners of English in Southern Tyrol"	"Its intent was to help the teacher better plan the English curriculum for the school year" "…the test helped students to better evaluate their own speaking competencies…" "No grades were awarded."	No specific cut score is mentioned, although contributor states that the test task and descriptors used for scoring is intended for learners at the B1 or B2 levels on *CEFR*. Contributor does not stipulate a competence level ("is sometimes true" "is often true" etc.) for descriptors

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

		she used for her scoring criteria.
Sakae Onoda, "An English Collocation Knowledge Test for College-level Learners and Pre- and In-service Teachers"	Contributor refers to the test as a "test/activity." "To raise learners' awareness of English collocations" To help learners automatize "such fundamental language units"	Cut score of 70%* as a pass/fail decision, stipulated by her institution for tests given for summative purposes. No method used, nonetheless she states: "Actively knowing 70% of the collocations is good news for students, and suggests that they can use collocations fairly well…"
Yesica Amaya, "La Historia De La Pola: An Achievement Test for Original Content-based Materials for Beginning Learners of Spanish"	"…to check learners' achievement on a lesson based on authentic materials I developed" "I decided to use these materials because one part of the course objectives stated that students should learn about different cultures where Spanish is spoken. However, the time spent in the classroom to work on culture was very limited…" "…the materials and the test was very limited since it was not part of the regular curriculum using in the course"	For the objectively scored part of the test the cut score was 14 out of 22 points (63.6%) which was determined using the contrasting groups method because she had scores from a master and a non-master group. For the performance test, the cut score was 10 out of 16 (66.66%) using the direct consensus method. She did not have master/non-master groups for this part of the test and so consulted with a colleague on what would constitute a C-score (a minimally competent learner) on her rating scales.

**Note.* Percentage scores reported in this book are assumed to be calculated by dividing a learner's score by the total possible score learners can get on a test. For instance on Sakae Onoda's contribution, a learner getting seven points on a ten point (ten item) test would get a score of 70% (7 points/items divided by 10 points/items = 70%).

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Note that three of the contributors used cut scores according to their purposes for giving the tests. These were essentially pass/fail decisions. It was not necessary for the contributors to create letter grades for the tests or to report the grades to their institutions. Yet they still wanted to know, for formative purposes, how their learners were doing. Yesica Amaya used the cut score to decide whether her learners had learned the materials. Her standard setting methods (contrasting groups and direct consensus) for setting the cut scores are defined below. Sakae Onoda used a cut score of 70% stipulated by her institution. She nonetheless comments that learners getting 70% of items right showed that whatever course content the learners were being tested on, they knew the content fairly well. This kind of thinking becomes important when doing standard setting, which is a family of methods for setting cut scores that involve consulting other instructors and experts on the content and the learner population.

Tests given for summative purposes. Most of the contributions fell into an in-between category, where contributors scored, graded, and reported grades from their tests (summative purposes), but also used the test results to give feedback to learners and to decide whether to review content (formative purposes). A few were used for purely summative purposes where the tests were given at the end of a semester, the scores were used to determine a course grade, and there was no opportunity to give feedback to learners. They are so noted in the table. See Table 7-2.

Table 7-2: ***Tests by contributors given for formative and summative purposes***

Contributor	Test purpose and test contribution to course grade	Cut score used and standard setting method used to set cut score
Dale Griffee, "A Vocabulary Quiz for ESL Learners at an Intensive Language School"	Formative: "I wanted to collect feedback…If the students all had failed, I would have to think about a review of the material." Summative: "…to calculate part of students' grades." 20% of course grade. "…it is unlikely that students could have	90-100 = A 80-90 = B 70-80 = C 60-70 = D Below 60 = F Contributor says he used a "traditional understanding of 90 = A, 80 = B, and so on" and purposefully made his test so the highest score would be 100. This would be termed a criterion-

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

	passed the course if they failed the test"	referenced cut score method (Gorsuch & Griffee, 2018, p. 251). This is not a standard setting method in the conventional sense.
Myles Grogan, "A Simple Speaking Test for an English-language University Communication Class"	Formative: To help learners "conceptualize their progress." "…to help students to be more comfortable with English and their interaction" Summative: To meet an institutional requirement for "a midterm and final for a course." Mid-term exam was 20% of the course grade; Final exam was 20% of the course grade. Learners must get a "C" in the course to graduate.	Contributor said he used an institutionally required grading percentage cut score system: Below 60 = D 60 69 = C 70 - 79 = B 80 89 = A 90 100 = S On one of his scales with a total of 13 points, he told learners the "target" was a "C" (9 points). On the second scale with a total of 10 points, he told learners the "target" was a 6, or 60% (A "C" on the university scale). No standard setting method is named, although his descriptors for a "C" on the 10 point scoring scale are detailed: "A score of 6 indicates a basic level of participation, with a small range of skills used…" The grade scale suggests a criterion-referenced cut score method.
Juliana Jandre and Vander Viana, "An English as a Foreign Language Test for Reading, Writing, and Cultural Diversity Awareness	Formative: "…the test was also instrumental in stimulating students to consider cultural diversity further."	The passing score of 6 out of 10 (60%) is stipulated by the school. There were 11 items on the test so a score of 60% would be approximately 6.5.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

for High School Students"	Summative: "To assess and support student learning." English is one of 13 disciplines learners must get at least a score of "6" out of a theorized "10" in order to graduate. This suggests a pass/fail decision. The test is 1/3 of three areas that are graded for the English course.	No standard setting method is named.
Ferit Kilickaya, "A Final Exam on Contextual English Grammar for Pre-Service Teachers of English"	Formative: "…to encourage learners to review the language structures that were practiced in context throughout the semester." Summative: "…to award part of a grade." The test counted for 60% of learners' course grades.	The test had 66 items but he weighted the scores so learners would get a score out of 100 points. Contributor did not decide cut scores for grades "…as my institution stipulated cut scores automatically when I was finished with entering learners' scores on the student affairs information system." No standard setting method is named.
Meredith Stevens and Meagan Kaiser, "Providing an Oral Summary of a Written Text as a Mid-semester and Final Test	Formative: "…to foster speaking skills." "…to lead learners to process English in its natural order." To get feedback from the mid-term exam "to help students practice collocations more effectively in preparation for the end of semester test."	The passing score of 60% is stipulated by the school. Learners can get a total possible score of 50 on either test. Thus, 30 points (60%) is a passing score. Specific scores for specific grades were not mentioned.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

	Summative: To award part of a course grade. The mid-term and final exams were worth 10% apiece toward learners' course grades.	No standard setting method is named.
Beatriz Garcia Glick, "An Oral VoiceThread for First-semester French Language Learners in a U.S. University"	Formative: "For students to practice speaking." To offer students feedback on their "written and pronunciation skills." Summative: To award part of a course grade. There are two tests worth 10% apiece.	Contributor stated that 60% is the pass/fail cut point. Learners can get a total of 5 points on the test, so 3 points (60%) is the cut point for the test. The contributor has descriptions for each point on the scale but no letter grades are associated with the points on the scale. No standard setting method is named.
Annis Shaver, "A Speaking Fluency Test for Intermediate-level German Using a Rubric Based on Grice's Conversational Maxims"	Summative: To assess whether learners achieved a course objective on "testing speaking proficiency." To award part of a course grade. The test was given four times and together the four tests were worth 10% of learners' grades.	Learners can get up to 20 points on each test. The author noted the following grading scale, which apparently represents percentages: 90 - 100 = A 80 - 90 = B 70 - 80 = C 60 - 70 = D Below 60 = F Thus a score of 12 out of 20 (60%) is the pass/fail cut score. The contributor does not name a standard setting method, but the scale she does give suggests a

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

		criterion-referenced cut score method.
Borbala Gaspar and Margherita Berti, "A Multiliteracies-oriented Project-based Assessment for Intermediate Foreign Language Italian Classes"	Formative: To encourage learners to develop "transcultural competence." To increase interest for learners by increasing their responsibility and control over their chosen content. To provide continuous feedback to learners. Summative: To award part of a course grade. The five part test accounts for 10% of learners' course grades.	Learners can get from 35 to 100 total points on the different parts of the test, which takes place throughout the semester. Contributors cite a 60% pass/fail cut score, meaning that on a 35 point test, 21 is the cut score. For a 50 point test, 30 is the cut score. For a 100 point test, 60 is the cut score. Contributors do not provide information on scales to assign a letter grade. No standard setting method is named.
Taichi Yamashita, "An End of Chapter Quiz and a Final Examination for Beginning-level Japanese Language Learners"	Formative: Give feedback to learners on their achievement of course outcomes. Summative: To award part of learners' course grades. Six quizzes are given in a semester which combined accounts for 20% of learners' course grades. The final exam is worth 20% of learners' course grade.	Contributor names 60% as a pass/fail cut score: "I primarily relied on my intuition that students who scored higher than 60% would be able to keep up with the subsequent chapters…" Contribution does not name a standard setting method.
Irina Drigalenko, "A Written and Oral Russian Achievement Test for Beginning	Formative: To offer feedback to learners.	The author noted the following grading scale "listed in the syllabus": 90 - 100 = A

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

College-level Learners"	To monitor learners' progress and "teach more effectively." Summative: To award part of learners' grades. The test accounts for 30% of learners' grades.	80 - 90 = B 70 - 80 = C 60 - 70 = D Below 60 = F Contributor designed her test to go up to 100 points. No cut score method is named but the scale above suggests a criterion-referenced cut score method.
Maria Martinez-Garcia, "A Final Project Performance Test for a Spanish Conversation Class at a Korean University"	Summative: To evaluate students' progress. To award part of learners' course grades. The test accounts for 40% of learners' course grades.	Contributor says her school sets cut scores for grades using a "relative grading system" meaning that "30 % of students are destined to receive a grade of C+ or lower. Further, a learner with 91 points out of 100 "may end up getting a C." Contributor designed her test to go up to 100 points. Even though she lists a letter grade for each point on the scale for each scoring criteria she must still follow the school's "relative grading system." No standard setting method is named. The description, however, suggests a norm-referenced cut score method (Gorsuch & Griffee, 2018, p. 248). Like criterion-referenced cut score method, this is not a standard setting method in the conventional sense.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

For the majority of tests in Table 7-2, no cut score methods are named. But another large majority of contributors use what appears to be a criterion-referenced cut score method, which is simply choosing even 10-point increments, descending from 100 to assign grades to. More on this method below.

Cut score. A cut score is a dividing score "above which is one status and below which is another status" (Gorsuch & Griffee, 2018, p. 317), such as learners passing the test at or above a score of 25 (out of, say, 50), but failing the test below that score. This would be a pass/fail decision. Zieky and Perie (2004, p. 2) add the idea that cut scores are used to categorize learners' scores on a test at several levels: "Cut scores are selected points on the scale of a test...student performance on a test may be classified into one of several categories such as basic, proficient or advanced on the basis of cut scores." In other words, decisions can be made with multiple cut scores on the scale of a test, such as A, B, C, etc. or whatever system of categories is used. Underpinning these cut score placement decisions is the assumption that learners' total scores on a test constitute a scale, and that learners' higher or lower scores on the scale constitutes more or less knowledge, or more or less ability. This is akin to contributors' scales used on scoring criteria such as 1 (low), 2, 3, 4, and 5 (high) to denote a difference in ability on "Student contributes enough information to the discussion" (see Amis Shaver's contribution), but then is expanded to the entire test, with all subtests and scoring criteria added up to a grand total. In the case of Amis Shaver's performance test, there are four scoring criteria with five possible points apiece. The highest possible score learners can get is 20. Her test score scale is from 5 – 20. Her cut scores are 18 (90%) for A, 16 (80%) for B, and so on.

A cut score method, then, is any process a teacher/tester uses to set cut scores along the scale of their test. As will be seen, there are different cut score methods, some of which do not involve standard setting, and some of which do. Cut score methods that do not involve standard setting employ a mechanical application of one or more cut scores on a total test scale with some assumption that learners fail at or slightly above the mid-point or mean score on the scale (see again Amis Shaver's contribution). This could be described as a top-down process. Cut score methods that do involve standard setting employ reasoned judgments of what learners know and can do based on individual test items and subtests, and test tasks, *before* any examination or construction of a total score scale for a test. This could be described as a bottom-up process.

Standard setting. Zieky and Perie (2004) whose report is highlighted in Further Reading, and which includes elements from a 1982 publication by

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Livingston and Zieky, shows roughly how long there has been sustained interest to seek meaning and descriptiveness behind learners' test scores, beyond simply comparing learners' scores to each other. It is no accident that the ACTFL *Guidelines* (American Council on the Teaching of Foreign Languages, 2012a) and *CEFR* (Council of Europe, 2001, 2018), with their descriptors and standards, both date from the 1980s (see Chapter Five). They represented efforts within second language education to define in mutually clear, non-numerical terms what learners could do at different levels of ability.

Key to standard setting is the idea of human judgment, as in Zieky and Perie (2004, p. 7): "All procedures for setting cut scores require the application of judgment." They emphasize the importance of employing qualified judges for standard setting, noting that the judges must be "educators who are subject-matter experts and who are familiar with students" and who are "aware of what students actually learn." Note the use of the term judges, plural, meaning that two or more judges are needed for standard setting. While there are many standard setting methods, the general logic or procedures for criterion-referenced tests and performance tests goes as described below in Table 7-3. Many elements are adapted from Cizek and Bunch (2007), and Zieky and Perie (2004).

Table 7-3: ***General logic and procedures for criterion-referenced tests and performance tests***

Criterion-referenced tests	Performance tests
1.) There is a determination as to what the cut scores are needed for (pass/fail, assignment of more than two categories as in letter grades, high or low stakes decisions).	1. There is a determination as to what the cut scores are needed for (pass/fail, assignment of more than two categories as in letter grades, high or low stakes decisions).
2.) Qualified judges are chosen.	2. Qualified judges are chosen.
3.) Judges examine subtests of a criterion-referenced test.	3. Judges examine the test task and the scoring criteria. Then they examine samples of learners' work (audio or video, or written work) that has been graded by someone else using the scoring criteria. *NOTE: The samples are presented in random order. No grades or levels or comments can*
4.) Judges independently give opinions as to what constitutes a minimally competent learner performance at a given level or levels for each subtest; this may go something like "6 out of 8	

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

items of this subtest probably indicates a learner of low 'B' ability" or "This subtest is very hard and a C- student would probably get only 2 out of 8 items correct."	*appear on the samples. Judges are not told the samples' scores, nor do they score the samples.* It is important there are multiple samples per level.
5.) Judges' determinations at a given level or levels are tallied for all the subtests (see Gorsuch & Griffee, 2018, p. 256 for a visual example).	4. Judges independently rank the samples from lowest to highest, then compare and discuss rankings. Some samples may be chunped together at the same rank.
6.) If there is consensus, then tallies of items correct out of each subtest are added up; that is the cut score for a given level.	5. Judges re-examine the test task and choose one or two learner samples for each level for which a cut score is required. The levels can be described in holistic, basic terms based on the scoring criteria, such as "accomplished the task in superior fashion" or "addresses the basics of the task" or "unable to accomplish the task."
7.) Evaluate the validity of the cut scores. For instance, are a lot of learners, thought to be high level based on other evidence, failing at cut scores below what they should?	6. Judges identify one minimally competent sample of work at each the two or three levels that are required.
	7. The scores previously (see item #3) given to the chosen minimally competent samples at each level are examined. These are the cut scores.
	8. Evaluate the validity of the cut scores. For instance, are a lot of learners, thought to be at a basic passing level on other measures, failing at cut scores that are lower than that level?

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

There are many standard setting procedures suitable for criterion-referenced tests (the Angoff Method, the Yes-No Method, Ebel's Method, etc.), but the general outline in Table 7-3 most closely resembles the direct consensus method used by Yesica Amaya (see her contribution). There are also many standard setting procedures for performance tests (the Bookmark Method, etc.) but the procedure outlined in Table 7-3 for performance tests most closely resembles the Body of Work Method. See both Gorsuch and Griffee (2018) and Zieky and Perie (2004) in the Further Reading section. See also the discussion below on the direct consensus method.

Note how both the criterion-referenced test and performance test standard setting procedures portrayed in Table 7-3 build upon judges' consideration of the actual test items and subtests and test tasks, and their suppositions of learners' interactions with them. Thus, learners are compared to the test content. Zieky and Perie (2004) note that the idea of a "borderline performance" (called "minimally competent" in Table 7-3) is key to setting cut scores, whatever method is used. Assuming three categories (three cut scores) of "basic" (a C), "proficient" (B), and "advanced" (A) are required, "distinguishing between the best performing student who is still basic and the worst performing student who is still proficient is necessary" (p. 8). The minimally competent learner at the bottom of a given level must be identified, hence steps 5, 6, and 7 for performance tests in Table 7-3. The minimally competent learners' performance is the anchor point for each level.

Cut score methods with standard setting used in this book. Two standard setting methods are mentioned by one contributor, Yesica Amaya, who used the contrasting groups method with her criterion-referenced test and the direct consensus method with the performance test section of her test (see Chapter Four).

Contrasting group method. This cut score method does not use judges but rather compares the test score distribution of two groups who take the test, a "master group" (learners who have studied the course content) and an "non-master group" (learners who have not studied the course content)(Horn, Ramos, Blumer, & Madaus, 2000, see Further Reading). The assumption is that the teacher/tester has carefully designed their test to capture course content (see Chapter Six on Practical Methods for Validating and Improving Tests). Learners who have studied the course content ought to score high on such a test. Learners who have not engaged in the course content ought to score low on the test. Yesica Amaya had two such groups. The highest possible score on the criterion-referenced portion of her test was 22. Thus her test score scale (for the criterion-referenced portion of her test) was 0 – 22. She only needed one cut score for a pass/fail decision, with

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

"pass" meaning the students did learn from her authentic materials. She administered her test to both master and non-master groups and scored them. She found that the lowest score in the master's group and the highest score in thenon-master's group overlapped at 14, which thus became her cut score. She also felt that master's group learners getting a 63.63% (14 divided by 22) were probably a model minimally competent C- student.

Direct consensus method. This was an unusual use of this method with a performance test. Typically the method is used with criterion-referenced test items or subtests. Nonetheless, given her short time frame and the resources she had at hand, it seems appropriate. Yesica Amaya explains she did not have a master's and non-master's group to use the contrasting groups method with this part of the test. She showed her test task and her scoring criteria to a colleague familiar with the students and with her materials. She stipulated that she was looking for a pass/fail decision. She then asked her colleague to pick out what scores on each of the three scoring criteria a minimally competent learner would likely get. With score scales from 1 (low) to 5 (high) on the three criteria, learners could get a maximum score of 15. Based on her colleague's comments she set her pass/fail cut score at 10 (66.66%). According to an e-mail from her (June 16, 2019): "For the cut score of 10, I would accept 4 ("organization"), 3 ("content"), and 3 ("grammar"). This is because the first criterion is the organization which is expected to have a logical sequence in the text. The other two are content and grammar which allow more flexibility since students may miss some events or have some grammar mistakes."

Default cut score methods without standard setting. Two default, top-down methods are given here, the criterion-referenced method and the norm-referenced method (see also Gorsuch & Griffee, 2018). "Default" simply refers to practices, in this case cut score methods, that are widely accepted and used without much thought. The two methods have some advantages. First, they are likely to pass muster at an institutional level. Teachers/testers using them will not raise eyebrows and supervisors are not likely to object to it. Second, teachers/testers may believe that the methods seem fair and objective, and thus easy to explain or defend the cut scores to stakeholders mostly notably learners and supervisors. Third, the cut score methods seems quick and easy to apply. Timeliness is always of concern to teachers/testers. But the methods also have serious disadvantages. The main disadvantage is that the cut scores are mechanically set without reference to the actual test content. In other words no human judgment on specific test items, subtests, performance test tasks, and scoring criteria has been used in relation to cut scores. There is no way to readily know, then, what particular

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

test scores mean. More details are given under each section on criterion-referenced and norm-referenced cut score methods below.

A second disadvantage is a lack of precision. When imposing cut scores mechanically onto a scale, no minimally competent level based on the test itself has been set for a pass/fail decision, or multiple cut score points for grades. This means that example responses of minimally competent learners at the cut scores are not available for examination. When teachers/testers begin scoring and then categorizing learners' tests, issues may come up with unexpected or alternate answers to test items or test tasks. Or, some learners' responses are simply hard to score and thus to categorize. It would be useful to have the responses of learners found to be minimally competent at various cut scores for decision-making in such cases. A final disadvantage is that a mechanical application of cut scores excludes the thinking of other educators and colleagues, who may have fresh insights on what constitutes expected knowledge or skill levels of learners who are grappling with course content. When a teacher/tester uses standard setting, the views of other informed judges are considered. A test task or test items may be harder or easier, or capture different knowledge and skills than the teacher/tester originally thought. There may be additional ways to think about learners' grasp of course content, and new ways to interpret their responses.

Criterion-referenced method. This is a commonly used method whereby teachers/testers calculate total scores on a test and then categorize the scores into four or five levels corresponding to letter grades of A, B, C, D, and F, or whatever categories are used by schools or education systems. Letter grades or categories are assigned using fixed score scale increments, such as ten-point increments, that represent percentages of test items answered correctly or appropriately. For instance, a pass/fail level is assumed to be 60% and then the grade categories are built up from there as in: 60 - 70 = D, 70 - 80 = C, 80 - 90 = B, and 90 - 100 = A. Thus the cut scores are 60 for a D, 70 for a C, and so on.

Many contributors used this method (Table 7-2), and some stated explicitly that they designed their tests or weight their test items so they add up to 100 points. As mentioned previously there is great utility to this, particularly when helping learners interpret the test results. Perhaps one reason for this seeming ease of interpretability is that the cut scores resemble a normal distribution (see Figure 3-1, Chapter Three) but with an added element of a logic of effort, for lack of a better term. The logic goes something like this: It is not enough that a learner should settle for being in the middle of the class, getting only 50% of the items correct (where the mean score is, and where the greatest number of learners ought to appear, if this were a norm-referenced test). Rather, they should strive to get at least

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

60% of the items correct, which is answering correctly beyond guesswork or chance. Further, even greater effort on learners' parts will result in being among the progressively fewer learners who get 70% correct, or 80% correct. While this logic does not fit the reality of score distribution of well-designed criterion-referenced tests (Figure 3-3, Chapter Three), it may seem to work as a means of motivating learners.

Nonetheless, applying pre-set, top-down categories obscures the meaning of learners' scores in terms of the actual test items, subtests, and test tasks. How do we know that learners who get 75% of the items correct should automatically get a C, just because the score is between 70 and 80%? What if this particular test has a new experimental subtest with new items on it, and the items are more difficult than we thought? Should not learners getting 75% correct with more difficult items get a higher grade than a C? Using a direct consensus method (a standard setting method) with the help of judges would alert teachers/testers to the fact that a particular subtest or test task is particularly challenging. For instance, Subtest 1 might have eight items which judges think are easy. They think a minimally competent C- student can get 6 items correct (75% is the cut score). But Subtest 2 (also eight items) might be of medium difficulty, where judges think a minimally competent C- student can get five items correct (62.5% is the cut score). And Subtest 3 (the new subtest, also eight items) might be judged to be very difficult, where judges think a minimally competent C- learner might get only two items correct (25% is the pass/fail cut score). Assuming the judges are in agreement, and the number of items judges associate with a minimally competent C- learner is added up to a total, then the cut score would be 13 out of 24 (54%), not 17 out of 24 (70.83%). In essence, because of the variable difficulty in items and subtests, the idea of equal intervals on a scale for cut scores no longer fits. Some items and subtests are harder. Clearly, as evidenced by the contributions to this book, teachers/tests put a lot of thought and time into designing their tests. It seems strange, then, that at this one juncture in the life of a test (deciding cut scores, see Figure 2-2 and Table 2-4, Chapter Two), a teacher/tester would hand over their subtle understanding of learners and the course content to a set of mechanical mathematical intervals.

Norm-referenced cut score method. This method is not appropriate for classroom tests. Nonetheless, it is worthwhile to describe the method as it is likely still used in education systems. Unfortunately, learners may have a lot of experience with their test scores being categorized using the method. And, as stated in Chapter Three, norm-referenced tests, and the idea of learners' test scores falling into a normal distribution, have a strong hold on

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

the popular imagination. Learners and administrators may expect this method to be used.

In one variation of this method, all learners' scores are totaled. Then the mean and the standard deviation for the class are calculated. This is easily done on a spreadsheet. The class mean score is the cut score for a C. One standard deviation below the mean is the cut score for a D. One standard deviation above the mean is the cut score for a B, and two standard deviations above the mean is the cut score for an A. Here is an example: On a criterion-referenced test with 50 possible points, the mean score for a class is 39. The standard deviation is 4.1. Thus the pass/fail cut score for a D is 39 – 4.1 = 34.9 (69.8% out of 50). The cut for a C = 39 (78%). The cut score for a B is 39 + 4.1 = 43.1 (86.22%). The cut score for an A is 39 + 8.2 (two standard deviations) = 47.2 (94.4%), which is very near the highest possible score learners can get (50). But, like the criterion-referenced cut score method, this method mechanically imposes categories upon a test scale without reference to the actual test items and test content. The categories of A, B, etc. are arbitrary.

With this method, even the intervals between cut scores are arbitrary. One year you may have a class that produces a mean score of 39 and a standard deviation of 4.1. Another year, you may have learners in your class who are more diverse with one or two learners who are very good (they had a year of study abroad in Japan, or wherever) or very poor (they came from poorly funded junior or senior high schools). You give the same test, and the mean is 35 out of 50. Because of the greater diversity in learners, your standard deviation is even larger at 6.5. Thus grade C level cut point is 35 (70% out of 50 items), your pass/fail D cut point is 35 – 6.5 = 28.5 (57%), your grade B cut point is 35 + 6.5 = 41.5 (83%), and your A grade cut point is 35 + 13 = 48 (96%). Why should this test, compared to the test given last year, be so much harder to get an A on, when a B or a C is much easier to get, with the larger intervals allowed for those two grades due to the larger standard deviation? Is that fair? Again, this type of cut score method is made without consideration of the actual test items or test tasks, and what performances learners ought to be able to achieve at a given level of study.

Chapter Summary

This chapter introduced practical methods for setting cut scores for both formative (non-grade assigning) and summative (grade assigning) purposes. The need to set cut scores was described in the context of tests being used by teachers/testers to make decisions about learners, including diagnosis and achievement. Key terms were defined, including cut scores, which is a

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

dividing score set on a test scale which marks whether learners pass or fail a test, or their test scores get categorized into grade levels, such as A, B, C, etc. Another key term was standard setting which is the process by which human judgments about cut scores are made with direct reference to test items, subtests, performance test tasks, and scoring criteria. Cut scores can be made with and without standard setting, and it was found from the contributions to this book that setting cut scores without standard setting was the norm, if only for this group of teachers/testers. Tables within the chapter specify contributors' cut score setting practices. Nonetheless, setting cut scores without standard setting entails substantial disadvantages. Mainly, without standard setting, cut score decisions are made without reference to the actual test items or test tasks. Thus, it is less possible to know what learners' scores actually mean on classroom tests that teachers make and use. Multiple explanations, examples, and resources for standard setting methods are given in the chapter in the hopes that more teachers/testers will try standard setting.

Further Reading

There are many sources available on setting cut scores and doing standard setting. Many are easily accessible, and written in clear, accessible language.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Thousand Oaks, CA: Sage.

Council of Europe (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment.* Retrieved December 1, 2018 from: https://rm.coe.int/16806667a2d

Gorsuch, G. & Griffee, D.T. (2018). *Second language testing for student evaluation and classroom research.* Charlotte, NC: Information Age Publishing.

Hom, C., Ramos, M., Blumer, I. & Madaus, G. (2000). *Cut scores: Results may vary.* The National Board on Educational Testing and Public Policy. Retrieved June 17, 2019 from: https://www.researchgate.net/publication/28799837_Cut_Scores_Results_May_Vary_NBETPP_Monographs_Volume_1_Number_1

Zieky, M. & Perie, M. (2004). *A primer on setting cut scores on tests of educational achievement.* Princeton, NJ: Educational Testing Service. Retrieved June 17, 2019 from: https://www.ets.org/Media/Research/pdf/Cut_Scores_Primer.pdf

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

CHAPTER EIGHT

PRACTICAL METHODS FOR USING TESTS FOR TEACHING AND LEARNING

GRETA GORSUCH

What this Chapter is About

This chapter is on how teachers/testers can use tests, and insights from learners' responses to tests, for teaching and learning. It was stated in the introduction to this book (Chapter One) that tests could, and should, be used for teaching and learning. There are several facets to this assertion, all of which play out in contributors' responses to the questionnaire shown in Chapter Two that gives this book flesh. One facet is how teachers/testers link their tests to their course content. In order to be used effectively for teaching, criterion-referenced tests and performance tests have to be related to course outcomes. In other words, a test has to have content validity (Chapter Six). As will be seen, several contributors consciously designed their tests to be similar to learners' classroom experiences, relying on the design and iterative use of the tests as tests, but also as classroom activities, to focus and motivate learners. Questionnaire items that probed this included: *What sources did you draw from for your test items? Test item ideas or content from a textbook? How did you get the ideas for what task learners had to do for the test? From tasks learners do in class?* (Table 2-2, Chapter Two).

Another facet is that tests give both teachers/testers and learners feedback on what learners know and do not know well, and what they can do and cannot do well. The feedback, in the form of test scores, subtest scores, test task scores, and teacher comments can be used by teachers to plan additional instruction or review sessions. The act of giving feedback can become the focus of a classroom activity or individual tutoring activity. Learners can use feedback to ask questions, clarify areas of uncertainty, plan revisions of their work, and then compare their new, revised work to earlier

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

feedback. Questionnaire items that probed this included: *Did you plan to give the scoring criteria to the learners for future self- or peer-assessment? Did you make another version of the scoring criteria for learners to use?* (Table 2-2). *Did you spend time going over the test in class? Did you teach learners to interpret their scores?* (Table 2-4). Yet another facet is that teachers/testers can use learners' test scores to plan future instruction. Questionnaire items that probed this included: *Did learners' test scores change your teaching? Did you re-teach content because learners didn't do well on your test?* (Table 2-4). A final facet is that tests can themselves be used for teaching and learning. Tests and taking tests promote memory retrieval in learners, a phenomenon observed in research exploring test effect. And, tests may themselves be made pedagogically worthwhile tests, meaning that teachers feel comfortable using them as classroom activities or tasks.

Examples of these uses of tests for all the facets suggested here will be given from contributors' comments. When juxtaposed with their actual tests and descriptions of their test procedures in Chapter Four, their commentaries offer practical applications of tests to teaching and learning. There are exciting developments in general education, and second language education that contribute to our practical, working understanding of how to more directly use tests for teaching and learning, and this chapter is but a small part of them. They offer a larger context of how tests and testing are being rethought as merely a means for awarding grades and categorizing learners (tests given for summative purposes), and more toward classroom-focused, non-institutional uses (tests given for formative purposes, see Chapter Seven). As with other content chapters in this book, there are key definitions given with examples from contributors' commentaries. The terms are content validity, feedback, test effect, and pedagogically worthwhile tests. Finally, there is a Chapter Summary, and a Further Reading section for sources in language education, general education, and psychology on tests and teaching and learning.

Definitions of Key Concepts

Content validity. As described in Chapter Six, content validity is the degree to which a test resembles what learners experience in class. Designing and using a test that has content validity is important for supporting test validity, which is the clarity, fairness, reliability, and appropriateness of a test and decisions made from scores from that test. But content validity is also important if tests are to be used for teaching and learning. For instance, teacher feedback from a test that is unconnected to course outcomes may

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

not support learners to reach course outcomes. As an example: You see your Spanish course as a way for learners to explore the culture of a South American country, and to improve their reading ability and persistence to read authentic texts from that country. There are course outcomes that specify this. As a result, you have learners reading authentic texts and working in groups to reflect on cultural content of texts. These activities comprise much of your classroom work. But then you are handed a required test from a supervisor that taps into knowledge of discrete grammar points presented in single sentences. The test does not tap into reading, but rather has a listening comprehension section of two men in an unspecified country asking and giving directions.

While the required test may be reliable and psychometrically sound, it is not related to what learners experience in class. It does not have much content validity. You can report learners' test scores to them and talk about what they might have done to improve their scores on the test, but your feedback has little to do with how learners can improve their reading and their ability to discern and discuss culture in authentic texts. Both the teacher and learners may find such feedback irrelevant or disconcerting. Even worse, such a test may strike learners as more "normal" to their experience and they might devalue the classroom instruction they have been getting.

Examples of contributors' treatment of content validity to support teaching and learning. Five contributors made obvious use of content validity to design and administer their tests. They may not have thought of their activity as promoting content validity. Rather, their thoughts, planning, and actions emerged from their current state of teacher theory (Figure 2-1, Chapter Two). They all depended on comparisons of the test to learners' classroom experiences which is in essence an understanding of content validity. The contributors also used the test tasks or test items themselves multiple times to help learners build up their mastery of the knowledge and skills the tests captured. Increasing learner experience with a test procedure and content, as well as repeated feedback to learners on the criteria of interest, would be a means of learning. In particular, repeated experience with a test may invoke test effect, which is a term from cognitive psychology, and is defined below. See Table 8-1.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 8-1: *Content validity in contributors' tests*

Contribution	Evidence of design based on content validity considerations
Myles Grogan, "A Simple Speaking Test for an English-language University Communication Class"	Content and procedure similarity: Class meetings throughout the course leading up to the test were taken up in practice and topics very similar to the test (Table 4-3, Chapter Four) Procedure similarity: Learners take the test twice during a semester, thus increasing learners' experience with real-time performance conditions. Repetitions and feedback: Learners get repeated feedback on criteria of interest. Notable quote: "In addition to an individual score, the group score gives students a simple metric with which to gauge their progress, and most students see a rapid gain over a semester." (see Figures 4-1 and 4-2, Chapter Four).
Meredith Stephens and Megan Kaiser "Providing an Oral Summary of a Written Text as a Mid-semester and Final Test"	Content and procedure similarity: Learners used texts for regular classroom activities from the same book with similar topics as the test texts. Content and procedure similarity: Learners read texts in class meetings, focusing on vocabulary, and then work in pairs to give oral summaries of the text using the highlighted vocabulary. This is similar to what learners do on the test. (see Tables 4-14 and 4-15, Chapter Four) Notable quote: "After learners work with the vocabulary test and hand it in, students are then asked to engage in a lengthy oral summary activity of the text." Procedure similarity: Learners took the test twice during a semester, thus increasing learners' experience with real-time performance conditions.
Annis Shaver, "A Speaking Fluency Test for Intermediate-level German Using a Rubric Based on Grice's Conversational Maxims"	Content and procedure similarity: Learners used similar discussion prompts for regular classroom activities as the test. (see Tables 4-22 and 4-23, Chapter Four) Notable quote: "The final discussion topic is set at least two days before the discussion for evaluation is scheduled. This helps to establish content validity (Moskal & Leydens, 2000)."

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

	Procedure similarity: Learners did six discussions that were the same as the test throughout the semester. Repetitions and feedback: Learners get repeated feedback on criteria (targets of instruction) of interest.
Taichi Yamashita, "An End of Chapter Quiz and a Final Examination for Beginning-level Japanese Language Learners"	Content similarity between classroom activity and textbook and test items: The content that inspired the items for both a unit test and a final exam were taken from daily classroom activities and textbook and lab program content. Content similarity between the unit tests and the final exam: The final exam was comprised of items sampled directly from the unit tests.
Irina Drigalenko, "A Written and Oral Russian Achievement Test for Beginning College-level Learners"	Content similarity between course outcomes and test content: Unit test content came from course outcomes. Notable quote: "Tests assist instructors in navigating their students from the testing point to next knowledge. Tests are also skill checkpoints on meeting the course objectives for both students and instructors." Content similarity between classroom activity and test: Teacher added test items to reflect new insights she had resulting from classroom instruction. She also added an option for learners to write a dialog because the task had worked well in class, where it was used for the first time. Notable quote: "For example, with Section II.1 on food and drinks, I started out started out with fewer items, but after reviewing the chapter before the test and seeing students' demonstration of successful verb usage in class, I decided to add more items to this section."

One contributor, Annis Shaver (Table 8-1) overtly invoked content validity in her commentary, along with a source (Moskal & Leydens, 2000), which can be found in full form in the Reference section. Regardless of whether contributors overtly thought of content validity while designing and using their tests, they acted in ways congruent with the concept. Their tests offered a platform for teaching and learning, helping learners build up an understanding of what knowledge and skills were needed to do well (Snow, Como, & Jackson, 1995).

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Feedback. The second key concept defined here is feedback. For this chapter, feedback will be defined as information from tests that teachers/testers notice or plan for, process using their teacher theory, and then use to guide teaching decisions. These decisions can be feedback to teachers where teachers alter their standing teaching plans in response to learners' performances on a classroom test or a performance test. These decisions can also be feedback to learners where teachers/testers offer learners test scores, written and/or verbal comments, interpretations of scores and comments, etc. as a means to extend and expand learning processes. See also the discussion in Chapter Seven on diagnostic decisions and achievement decisions made with tests.

Feedback is part of a larger concept gaining acceptance in second language education, that of tests as a means of learning. This larger movement is known by various names, such as classroom assessment (Cheng, Rogers, & Huiqin, 2004; Wang, 2017), teacher-based or teacher assessment (Davison, 2004; Rea-Dickins, 2004), diagnostic assessment (Doe, 2015), formative assessment (Brumen, Zupancic, Aguero, & Alonso-Belmonte, 2018; Doe, 2015; Fulcher, 2010b), assessment for learning (Colby-Kelly & Turner, 2007), and tests given for formative purposes (see Chapter Seven). Classroom assessment is an overarching concept which is thought to be a fluid, contingent process with three components: Teachers collecting information on learners (such as test scores or observations), teachers making judgments (decisions) about the information, and then teachers using the information in some way (Wang, 2017). This section deals with how the information is used, in this case, in the form of feedback.

Feedback to teachers. Teachers may use feedback from tests to alter their standing teaching plans either in the short term for a current course, or in the long term for future courses. Teachers may also use their insights from test results to assist in tutoring learners who approach them individually. And, teachers/testers may use learners' responses to tests to change and improve future versions of their tests. Contributors' comments suggest their active engagement in all four. Information that teachers gain about learners from tests provides food for thought for teachers' teacher theory, which is fluid, efficient, and goal oriented (Figure 2-1, Chapter Two). For instance, finding that learners did not do well on particular test items, sub-tests, or test tasks teachers/testers would then need to solve this problem ("Problem" in Figure 2-1), balancing this need with time and other resources available ("Current teaching," Figure 2-1).

Contributors' short- and long-term teaching responses. Contributors to the book said they engaged in both short- and long-term changes in their teaching. In the short-term, interestingly, no contributors to this book

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

actually changed standing teaching plans in response to feedback from tests. There were no emergencies, or as Irina Drigalenko put it, no one "gave up or fainted" taking the test. No learners did so badly or so well that it was alarming. Rather, contributors addressed test results in the classroom in already-planned sessions as soon as scoring was done. Three things stand out: First, teachers/testers were concerned with timeliness. They wanted to give test outcomes to learners very soon after the test, and then process the outcomes in some way in class. Second, contributors varied greatly in how they processed outcomes with learners; ranging from short teacher-fronted lectures to longer, more interactive classroom activities, to activities that actually spanned multiple class sessions. For some contributors who used interactive treatments of test review, learners could tailor their questions to their own concerns and areas of uncertainty. Third and finally, contributors varied greatly in the level of specificity of their descriptions of what they did. Some contributors were general in their descriptions of post-test sessions. Yet others were detailed. See Table 8-2.

Table 8-2: ***Contributors' short-term teaching responses to test results***

Contribution	Timing and duration	Teaching responses
Dale Griffee, "A Vocabulary Quiz for ESL Learners at an Intensive Language School"	Most of the following class meeting (30-40 minutes, within 24 hours).	-Tests were handed back, then re-collected at end of class. -Going over the test was a classroom activity.
Juliana Jandre and Vander Viana, "An English as a Foreign Language Test for Reading, Writing, and Cultural Diversity Awareness for High School Students"	The following class meeting, unknown interval, unknown duration.	-Tests were handed back, learners got to keep them. -Teacher identified trouble points in learners' responses. -Teacher commented on how answers could be improved. -Teacher elicited answers from learners. -Teacher elicited learner questions about specific problems they had on the test.
Sakae Onoda, "An English Collocation Knowledge Test for College-level Learners and Pre-	The following class meeting after each time this type of test was given, unknown duration.	-Tests were handed back, then re-collected after class. -Teacher focused on test items that learners got wrong. -Teacher explained answers and gave key definitions.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

and In-service Teachers"		-Teacher summarizes verb collocations found in class texts and materials at the end of each class if time allows.
Meredith Stephens and Megan Kaiser "Providing an Oral Summary of a Written Text as a Mid-semester and Final Test"	Scores reported individually to learners one week after the test. Further oral feedback given in class one week after the test. New content and "old" content retaught in multiple following lessons.	-Re-taught content in the following weeks of the course, including how to choose key collocations and words for making summaries from class texts used that day, and giving learners practice to remember collocations rather than directly translating from Japanese. -Emphasized connection between having a wide vocabulary and communicating quickly. -Surprised at how quiet learners in class did well on the test and how talkative learners in class did not necessarily do well on the test, teachers minimized teacher talk in subsequent classes and offered more learner talk time.
Beatriz Garcia Glick, "An Oral VoiceThread for First-semester French Language Learners in a U.S. University"	Scores reported to learners one week after the test. Unknown duration.	-Teacher addressed common mistakes and specific concerns of students. -Spent more time in following classes on present tense.
Borbala Gaspar and Margherita Berti, "A Multiliteracies-oriented, Project-based Assessment for Intermediate Foreign Language Italian Classes"	Learner scores and teacher feedback on successive parts of the tests retrieved by learners on an electronic course interface. Scores and feedback posted within ten days after each part of the test was completed.	-Teachers gave more in-class workshops to help learners comprehend online L2 content they wanted to use in their test/projects. (Much of their feedback and responses to learner feedback responses was done in private tutorial fashion, see Tables 8-4 and 8-5 below.)

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Taichi Yamashita, "An End of Chapter Quiz and a Final Examination for Beginning-level Japanese Language Learners"	Most of the following class meeting.	-Teacher says his response was largely teacher fronted. -Teacher explained answers to each item on the test. -Teacher played audio files with listening passages to help learners confirm their answers.
Irine Drigalenko, "A Written and Oral Russian Achievement Test for Beginning College-level Learners"	Most of a class meeting two to three days after the test.	-Tests handed back, students kept the tests and had the option to self-correct the exams and have the teacher comment on their self-corrections. -Teacher focused on common errors. -Teacher elicited answers from learners.

Many contributors used feedback from their tests to change their teaching in the long-term for future courses. For some contributors, learners' responses to their tests comprised a gold mine of awareness upon which to plan sometimes very specific teaching plans. See Table 8-3.

Table 8-3: ***Contributors' long-term teaching responses to test results***

Contribution	Teaching responses
Kai-Ying Hsu, "A Chinese Achievement Test for Intermediate College-level Learners"	-Teacher believes vocabulary needs more practice in class. -Teacher wishes to have learners listen to more native speakers speaking using audio files. -Teacher wants to have learners deconstruct listening audio files and try listening to them repeatedly in class.
Dale Griffee, "A Vocabulary Quiz for ESL Learners at an Intensive Language School"	-Teacher wants to make several test tests of the kind he contributed to this book and use them as classroom activities.
Myles Grogan, "A Simple Speaking Test for an English-language University Communication Class"	-Teacher plans to emphasize more practical communication and conversation in class. -Teacher wants to consider having learners record and transcribe their own classroom talks.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Juliana Jandre and Vander Viana, "An English as a Foreign Language Test for Reading, Writing, and Cultural Diversity Awareness for High School Students"	-Teachers want to more emphasize the legitimacy of non-American or British varieties of English.
Ferit Kiliçkaya, "A Final Exam on Contextual English Grammar for Pre-service Teachers of English"	-Teacher plans to pay greater attention to specific structures learners did poorly on on his test. -Teacher wishes to re-order the structures in which he teaches them.
Gisela Mayr, "A Speaking Skills Test for High School Learners of English in Southern Tyrol"	-Teacher wants to emphasize speaking on a daily basis in future classes. -Teacher plans to have learners engage in more student to student interaction. -Teacher hopes to choose for current content and texts for learners on which to base learner speaking tasks. -Teacher wants to use more authentic texts in class.
Borbala Gaspar and Margherita Berti, "A Multiliteracies-oriented, Project-based Assessment for Intermediate Foreign Language Italian Classes"	-Teachers plan to teach learners to use a 3-2-1 summary activity do learners can better evaluate authentic L2 texts and websites they use for their test. -Teachers want to have more in-class practice to find information and make sense of resources found in target language web pages.
Maria Martinez-Garcia, "A Final Project Performance Test for a Spanish Conversation Class at a Korean University"	-Teacher wants to include more group activities in class where learners solve some sort of problem in groups.

Contributors' responses for tutoring purposes. A few contributors mentioned specifically that information they captured from their tests and the structure of the tests themselves assisted them to address tutoring situations. A special case was Gaspar's and Berti's test, which was based on the middle-level theory of multiple literacies (see Chapter Five). With such courses and tests, individual tutoring and a constant cycle of learner work → instructor feedback → revised learner work is theorized as a main means of student learning. See Table 8-4.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 8-4: ***Contributors' responses for tutoring sessions***

Contribution	Tutoring responses
Myles Gregan, "A Simple Speaking Test for an English-language University Communication Class"	-Learners come in for coaching individually; The test with its scoring criteria give structure for individual work with learners. -Learners are largely concerned that some people in their class and test groups talk too much or too little. -Teacher uses test scoring criteria and test task to contextualize advice to learners in using nonverbal interaction signals.
Sakae Onoda, "An English Collocation Knowledge Test for College-level Learners and Pre- and In-service Teachers"	-Based on what she learned about learners' knowledge of collocations, teacher prepared a suite of suggestions for learners in tutorial sessions including websites, listening and reading materials, and dedicated software.
Borbala Gaspar and Margherita Berti, "A Multiliteracies-oriented, Project-based Assessment for Intermediate Foreign Language Italian Classes"	-The test as structured requires multiple drafts of several phases to the test over time. Tutoring took place online in confidential, synchronous messaging. -Teachers creating scoring criteria for the different phases, creating specific structures and content used in tutoring (see Tables 4-25, 4-26, Chapter Four).

Contributors' responses to feedback to revise their tests. Some contributors used learners' responses and their own observations during testing procedures to consider revisions to their tests for future use. This accords with Downing's (2006) guidelines for test design and the "Using test scores" component of the "Life of a Classroom Test" model (Figure 2-2, Chapter Two), which is based on Downing's observations. In the model, an arrow goes from "Test results" directly back to the "Test planning and writing" component. See Table 8-5.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Table 8-5: ***Contributors' responses for test revisions***

Kai Ying Hsu, "A Chinese Achievement Test for Intermediate College-level Learners" Contributor wishes to improve the sound quality of the video and audio files for her test.
Myles Grogan, "A Simple Speaking Test for an English-language University Communication Class" Contributor wants to better assess learners' declarative knowledge of communication strategies needed for the test task. This may involve using discourse completion tasks or learner transcription of their test performances. Doing this may make more clear the current test task for learners.
Juliana Jandre and Vander Viana, "An English as a Foreign Language Test for Reading, Writing, and Cultural Diversity Awareness for High School Students" Test could be made longer, as learners finished the current test well ahead of time. Contributors also want to use entire original texts as prompts, rather than excerpts. Contributors also want to mark more clearly on the tests how much each item or task is worth so learners can make their own decisions about how they wish to treat items, and about the order in which they answer items.
Ferit Kiliçkaya, "A Final Exam on Contextual English Grammar for Pre-service Teachers of English" Plans to shift away from dichotomous (two answer) response items to items with more responses so learners do not just get items correct by chance. Contributor also wants to increase the amount of time allowed for the test, as some learners ran out of time.
Meredith Stevens and Meagan Kaiser, "Providing an Oral Summary of a Written Text as a Mid-semester and Final Test Reflecting on their observations of the mid-term exam test procedure, contributors revised the test task, asking learners to use collocations rather than single-word vocabulary items for their oral summaries (see Table 4-15, Chapter Four). They also re-weighted their scoring criteria to account more for learners' use of collocations in their oral summaries relative to other components of the scoring criteria (see Table 4-19, Chapter Four).

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Taichi Yamashita, "An End of Chapter Quiz and a Final Examination for Beginning-level Japanese Language Learners" Contributor wants to arrange the listening subtest items to appear in the same order in which they appear in the audio file that goes with the test. He thinks some learners who got high overall grades on the test got some listening subtest items wrong because the items appeared in a different order than suggested by the audio file.
Maria Martinez-Garcia, "A Final Project Performance Test for a Spanish Conversation Class at a Korean University" Contributor wishes to expand her test to be a mid-term and a final exam.

Contributors vary in their area of concern with their test revisions. For instance, Myles Grogan wished to measure learners' knowledge in addition to what he already captures in his test. Others wished to improve or revise the texts or test items that learners respond to. All contributors were relatively specific in what they want to revise. It is interesting to consider whether the tests in and of themselves create a specific focus for attention for teachers as well as for learners. As will be discussed below in the section on test effect, interacting with test items, subtests, and test tasks, and having experiences taking a test, creates a focus for learners which help to draw attention to a test and to retrieve memories relevant to the test. It is also notable that Stephens and Kaiser actually did their revisions to the weighting of scoring criteria in mid-semester.

Feedback to learners. Feedback has a "powerful but variable influence on FL learning by giving learners the opportunity to enhance their metalinguistic awareness" (Brumen et al, 2018, p. 217). In other words, feedback can be used by learners to explore their own learning processes—what they notice, what they know and do not know, and what they want to do about it (see also Doe, 2015). Brumen et al go so far as to assert that timely and effective feedback are "a fundamental feature of good teaching practice" (2018, p. 217). Thus, feedback must be given "at appropriate points in time" in a stress-free way (2018, p. 217). Others comment on how to make feedback effective, including Fulcher (2010b) who states that effective feedback has to: 1.) Offer learners information on their current state of ability, and 2.) Help learners identify the contrasts between what they can currently do with some future state of what they want to be able to do. Finally, 3.) Effective feedback must include guidance on how to get from where learners are now, to where they wish they wish to go (Fulcher, 2010b).

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Contributors to this book had many ways to give learners feedback, some methodologically simple and others complex, and also internally coherent and extended. In addition to the features of effective feedback suggested by Brumen et al (2018) and Fulcher (2010b), contributors' responses suggested additional criteria to consider for effective feedback, including: 4.) Using feedback iteratively (repeatedly) to increase learners' experience with tests, test tasks, and scoring criteria; 5.) Ensuring learners understand the scoring criteria for a performance test; and 6.) Allowing re-testing or revision and resubmission opportunities, if possible and/or appropriate. See Table 8-6.

Table 8-6: ***Feedback to learners described by contributors to this book***

Contribution	Feedback to learners
Dale Griffee, "A Vocabulary Quiz for ESL Learners at an Intensive Language School" Criterion-referenced test	General description: Graded tests handed back to learners. A test review session is held where teacher mentions trouble points, goes over items, elicits answers from learners. Learners know where they are: Teacher designs his test to add up to 100, to help interpretability of scores.
Myles Grogan, "A Simple Speaking Test for an English-language University Communication Class" Performance test	General description: Learners get oral and written feedback on a score sheet immediately after each test. There is a mid-term test and a final test. Learners know where they are: Teacher offers brief, targeted feedback during test. Learners get verbal feedback and a marked score sheet immediately after test. Learners contrast where they are and where they want to be: Learners self-assess in classroom tasks that are similar to test tasks, and peer-assess another group in the class. Iterative feedback: There is a mid-term test and a final test with similar test tasks and the same scoring criteria. Learners get similar feedback on classroom tasks. Scoring criteria comprehension: Learners practice scoring system for number of utterances in class previous to test. Learners shown an example marked up score sheet previous to test. Retesting: Learners can re-test if group dynamics in the test break down. Teacher gives judicious coaching relevant to scoring criteria and learners re-test later in the test period.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Juliana Jandre and Vander Viana, "An English as a Foreign Language Test for Reading, Writing, and Cultural Diversity Awareness for High School Students" Criterion-referenced test	General description: Graded tests handed back to learners to keep. A test review session was held soon after the test. Feedback was written and oral. Learners know where they are: Teachers designed test to add up to ten points, which is the same scoring system used at the school for other courses. A pass/fail cut score of 6 is used throughout the school. Guidance: Written feedback to individuals and post-test review session with the whole group includes guidance on how to improve. Scoring criteria comprehension: Learners asked questions and got answers on the scoring criteria in the post-test review session.
Ferit Kilickaya, "A Final Exam on Contextual English Grammar for Pre-service Teachers of English" Criterion-referenced test	General description: Learners got their test scores online on a student affairs system at the university. Learners requesting an individual meeting after the test could do so face to face or on e-mail. Learners know where they are: Test designed so total score was 100. Learners contrast where they are and where they want to be: Teacher sent the test and the answer key to learners after the test by e-mail and social media.
Gisela Mayr, "A Speaking Skills Test for High School Learners of English in Southern Tyrol" Performance test	General description: Learners get their score on a marked score sheet, and are given oral feedback immediately after the test. Learners know where they are: Written and verbal feedback is given immediately after the test when learners' impressions were still fresh. This was the only test in this book given *before* instruction the teacher wanted learners to get experience with this type of test, which they had never experienced before, so they could self-evaluate. Scoring criteria comprehension: Learners got the scoring criteria to read before test. Teacher answered questions from learners. Scoring criteria were in learners' L1.
Sakae Onoda, "An English Collocation Knowledge Test for College-level Learners and Pre- and In-service Teachers" Criterion-referenced test	General description: After the test, learners correct each other's tests in pairs, there is a review session where correct answers are elicited and also given, and the tests are handed in to the teacher for final checking. Teacher hands back the graded and checked tests with written feedback and a second review session. Timely, extended feedback is emphasized so learners can "modify any imperfect knowledge as soon as possible." Learners know where they are: Test designed to add up to ten points with the school-wide assumption that

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

	seven points is 70% and a pass/fail cut point. Learners correct each other's tests and participate in a review session where teacher elicits their answers. Learners are taught how to give partial credit if partners use incorrect verb forms. Learners calculate the total scores. When teacher gives back the re-checked tests again in a few days, she explains the class test mean and standard deviation. Teacher believes pair scoring helps learners think "about their own likely score." Learners contrast where they are and where they want to be: In second review session, teacher answers learners' questions about correct answers. Guidance: In group and individual feedback sessions, teacher gives tips and offers resources on how to improve.
Meredith Stephens and Megan Kaiser "Providing an Oral Summary of a Written Text as a Mid-semester and Final Test" Performance test	General description: Learners get written feedback on a score sheet with scoring criteria one week after the test. There is a mid-term test and a final test. Learners know where they are: Teachers focused on writing scoring criteria that offered positive evidence and not negative feedback focusing on lack. Iterations: Learners did similar test tasks for course work. Learners did the same test task for both a mid-term and a final test, using similar scoring criteria. Scoring criteria comprehension: Teachers sent out scoring criteria to learners by e-mail two weeks before the test.
Beatriz Garcia Glick, "An Oral VoiceThread for First-Semester French Language Learners in a U.S. University" Performance test	General description: Learners submitted their recordings online, and the teacher responded with individualized written and spoken (recorded) feedback online. Teacher also reviewed the test verbally in class. Learners know where they are: Learners get written feedback in the form of scoring criteria with simple, explicit descriptors. Her written and oral feedback related directly to the scoring criteria. Teacher commented on common, major mistakes made by learners in a review session. Scoring criteria comprehension: Scoring criteria are in the course syllabus. Teacher believes the scoring criteria are self-explanatory. Re-testing allowed: Learners can submit a second recording for full credit after getting feedback on their first recording.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Annis Shaver, "A Speaking Fluency Test for Intermediate-level German Using a Rubric Based on Grice's Conversational Maxims" Performance test	General description: Learners get scores on a score sheet and additional written comments and suggestions after the test, uploaded to a Moodle course site for individual student review. Learners know where they are: Teacher writes detailed comments on learners' score sheets. Learners contrast where they are and where they want to be: Teacher sometimes sets individual interviews with learners to give guidance on how to improve their participation in the discussion [the test] and to explain her expectations. Iterations: The test is administered six times in a semester using very similar test tasks and the same scoring criteria. Scoring criteria comprehension: Teacher believes the scoring criteria are straightforward.
Borbala Gaspar and Margherita Berti, "A Multiliteracies-oriented, Project-based Assessment for Intermediate Foreign Language Italian Classes" Performance test	General description: This is a series of five performance tests that comprise a project. Learners get feedback on each test on an online platform. Two-way feedback between the teachers and learners is practiced. Learners know where they are: Feedback is two-way and dialogic between teacher and learners (online and in class), and learners and learners (in class). Learner questions are encouraged. Learners contrast where they are and where they want to be: Learners do two of the tests twice, and each of the repeated tests gets its own grade. Guidance: Learners get guidance from teachers as part of the dialogic feedback. Teachers are seen as a giver of assistance to learners. Iterations: Two performance tests use one set of scoring criteria, two performance tests uses a second set of scoring criteria, and the final performance test uses a third set of scoring criteria. Thus learners get repeated experience with four out of five of the tests. Scoring criteria comprehension: Learners get practice reading and using the scoring criteria during class workshops designed to help them find materials and sources for their tests/projects.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Taichi Yamashita, "An End of Chapter Quiz and a Final Examination for Beginning-level Japanese Language Learners" Criterion-referenced test	General description: Learned got scored and graded tests back after the test, and could keep them. Teacher did a formal review of the test in class. Learners know where they are: Teacher told learners the class average score. Teacher verbally explained each test item, and also emphasized common mistakes and problems learners had on the test.
Irina Drigalenko, "A Written and Oral Russian Achievement Test for Beginning College-level Learners" Criterion-referenced test and Performance test	General description: Learners get scored and graded tests back after test, and get to keep them. There is a post-test review session done in-class. Learners know where they are: Teacher gives both written and oral feedback. Teacher believes written feedback should be more precise and detailed. Learners contrast where they are and where they want to be: Learners got to keep unit tests to prepare for the final exam. Post-test review session had learners self-correcting their answers and providing answers during the discussion. Teacher allowed learners to self-correct their own tests and hand them in for further feedback from teacher. Teacher's stated goal is to help learners self-investigate their own errors and to engage in self-directed improvement. Guidance: Teacher gave learners a pre-test handout to guide study for the test based on errors learners made on previous homework assignments. Scoring criteria comprehension: Learners got the scoring criteria to study for the performance tests before the test day.

The categories used in the right hand ("Feedback to learners") column in Table 8-6 emerge from the six points suggested for effective feedback given above. The categories have permeable boundaries, including "Learners know where they are" and "Learners contrast where they are and where they want to be." This suggests that a few simple methodological changes a teacher/tester makes can turn a feedback session from awareness raising ("Learners know where they are") to helping learners contrast their current state with what is possible, a practice supported by research in cognitive psychology (see for example Graham & Weiner, 1995). For instance, Irina Drigalenko adds learner self-correction in and out of class,

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

with feedback on learners' revised work to the palette of feedback processes offered to learners.

Test effect. The third key concept for the chapter is test effect. Test effect "is the idea that taking a test, in and of itself, is conducive to learning" (Gorsuch & Griffee, 2018, p. 270). In multiple experiments, evidence has emerged that taking a test may help learners learn and remember more than studying or re-reading notes or other materials (Butler & Roediger, 2007; Roediger, Agarwal, McDaniel, & McDermott, 2011; Roediger & Karpicke, 2006). The concept comes from cognitive psychology. It is posited by scholars working in this tradition that test effect works through learners' memory mechanisms, and engages learners in recalling, reasoning, planning, reducing anxiety, and spacing out study. One experiment done by a cognitive psychologist (Leeming, 2002) offers an example of how test effect may work. Leeming had two groups: College-level psychology learners who were tested in the usual way with four regular exams during the semester (mid-terms and a final), interspersed with regular lectures. The second group had short, ungraded quizzes at each class meeting using the same content and test item types planned for the learners for the final exam for the course. The instructor repeatedly reminded learners the quizzes were not graded and would not count toward their final grades. The second group taking the frequent quizzes got on average half a grade higher on the course than the first group. Learners in the second group also reported studying more. The description suggests that learners' memory processes were activated and that they engaged their metacognition, in other words their ability to form plans and self-direct their attention and energy.

Test effect shows up in a few second language studies, although it is not called that. Hawkes (2012) pointed out a backlash against task-based language teaching where teachers found that learners working in groups and doing meaning-focused tasks did not necessarily use the grammatical forms the teachers wanted learners to focus on (2012, p. 327). Hawkes wanted to find a way to make a task more form-focused and so proposed the following task methodology: 1.) Learners do a pre-task, 2.) Learners do the main task, 3.) Learners immediately engage in "form-focus…where predetermined structures were highlighted and practiced," (p. 327) and 4.) Learners do the exact main task (#2) once more. Hawkes found in transcribed audio recordings that when learners did the repeated task "more attention was being placed on form" (p. 334). Learners were using the grammatical forms of interest more, and more accurately. Hawkes attributes this to shifting learners' limited attention to form in step #3. But test effect may have also played a role in learners' improvement. The tasks in steps #2 and #4 were not intended as tests, and were not graded, presumably. But the tasks

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

comprised domain-specific experiences (Zimmerman & Schunk, 2008), which is the essence of a test. Doing the tasks twice likely invoked learners' general psychological processes of recalling, reasoning, and planning. Hawkes notes that learners could do the task re-do using any notes they took during the form focus session (step #3). Would not the notes, and use of them, be taken as additional evidence of recalling and planning? See also Wang (2008) for a study on task-based language teaching in which the terms "pre-task planning" and "rehearsal" (p. 84) may be partial proxies for test effect, although again, this concept is not mentioned and the author may think of his terms using a different theoretical frame.

Contributors Grogan, Stephens and Kaiser, Shaver, and Gaspar and Berti, among others, all gave learners repeated experiences with their tests, test tasks, and scoring criteria. While no contributor mentions test effect, something like this concept must be playing a role in contributors' designing and using their tests. See also Table 8-6 for contributors who used "Iterations" as part of their feedback practices.

Pedagogically worthwhile tests. The fourth and final concept defined here is pedagogically worthwhile tests. Briefly, pedagogically worthwhile tests are criterion-referenced tests and performance tests that teachers would feel comfortable using as teaching materials, and as the focus for actual teaching and learning in the classroom (Gorsuch & Griffee, 2018). This may bring up terrible images of teachers drilling learners with multiple choice test items on truly random and unpredictable norm-referenced test content (see Chapter Three). Presumably, some second language teachers/testers are comfortable with such tests and may be happy using them to teach. But the concept of pedagogically worthwhile tests discussed here is more connected to general shifts in educational testing in the U.S. and elsewhere toward a greater appreciation of formative assessment (also called classroom assessment, tests for formative purposes, testing for learning, etc.). Formative assessment features eliciting learners' prior knowledge and engaging their metacognition and ability to self-monitor (Shepard, 2000a). A cornerstone of formative assessment is teachers giving feedback from tests to learners as a means of learning (Shepard, 2000a, 2000b, 2005; see also the discussion in this chapter on feedback).

In particular Shepard called for a general improvement in "the content and character" of tests themselves (2000a, p. 1) so that tests could be used "interchangeably" as teaching materials and as tests (p. 38). This means moving away from testing declarative knowledge and facts through objective test items, to open-ended tests which "develops a community of practice where it is customary for students to review and question what they already believe" (p. 20). She suggested features of such tests: 1.) Tests

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

should be open-ended and extended, and avoid having learners simply choose a single answer; 2.) Tests should elicit learners' prior knowledge; 3.) Tests should provide feedback; 4.) Tests should engage learners in self-evaluation; 5.) Tests should encourage learner reasoning and inquiry; 6.) Tests should include ways for learners to communicate their evolving understandings; and 7.) Tests should build in practice needed to complete the test or task requirements (Shepard, 2000a, 2000b, 2005).

It is suggested here that pedagogically worthwhile tests are not simply a matter of form or how they look, even though Shepard names "reflective journals, oral presentations, work samples, projects, and portfolios" (2000a, p. 43) as examples of tests useful for classroom assessment. Rather, there is a methodological aspect to the idea of pedagogically worthwhile tests. In other words, what makes a test worthwhile as a focus of teaching and learning is what you do with a test in class. For instance, Hurt (2019, p. 1) suggests building an ordinary objectively scored grammar test from weekly reviews in class, working together with students each week to select test content and items, and also to decide "what would qualify as mastering the lesson." Hurt further suggests doing the weekly test building tasks in a student circle so the teacher is seen as a facilitator, and to build in learners communicating their suggestions for the test to the class. Beare (2019) works from an objectively scored test to begin with, suggesting the test items be used as reviews leading up to the test. Reviews would include learners writing their answers on the board, other learners indicating whether they had the same answer, having students say why they chose a particular answer and why "each incorrect answer is incorrect" (p. 4). Thus, even when objectively scored tests are used, they can be made pedagogically worthwhile by extending learners' engagement with them and turning them into a subject of inquiry. Nearly every contributor to this book offers pedagogically worthwhile tests, either through test design, or by how tests are used for teaching and learning in class. See Tables 8-1, 8-2, 8-3, 8-4, and 8-6, and also contributors' commentaries in Chapter Four.

Chapter Summary

This final content chapter addressed the idea of using tests themselves for teaching and learning. Using tests for teaching and learning was considered through the key concepts of content validity, feedback, test effect, and pedagogically worthwhile tests. Feedback was explored as feedback to teachers and feedback to learners. Feedback to teachers was defined as information from tests that could propel teacher planning of short-term and long-term changes to teaching, and as a resource for tutoring and for

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

revising future versions of the test. Contributors to the book engaged in all four uses of feedback to teachers. In terms of feedback to learners, effective feedback was characterized as offering learners information on their current states of ability, helping learners see the gap between where they are and where they want to be, giving guidance, using feedback iteratively, and ensuring learners understand scoring criteria, among other features. Contributors to this book had many ways to give learners effective feedback, some of them amounting to complex and internally coherent methodologies (see Table 8-5). Finally, the key concepts of test effect and pedagogically worthwhile tests were explored. Test effect refers to tests being used by teachers as a focus of attention to engage learners in learning processes including recalling, reasoning, and planning. Small, ungraded, frequent tests may provide learners the domain-specific experiences they need to do well on a longer and similar graded test (tests given for summative purposes). Pedagogically worthwhile tests, tests that can used interchangeably as teaching materials and tests, were set in a context of general movements in educational testing, that of formative assessment, assessment for learning, and classroom assessment. In this context, tests are seen as opportunities for learners to engage in self-evaluation, reasoning, and inquiry through cycles of extended work, feedback, and discussion.

Further Reading

The following are resources on using tests for teaching and learning. They are from second language education, and also general educational testing and cognitive psychology.

Content Validity

Siddiek, A.G. (2010). The impact of test content validity on language teaching and learning. *Asian Social Science, 6*(12), 133-143. Retrieved May 25, 2019 from: https://files.eric.ed.gov/fulltext/ED574721.pdf

General Testing Books or Resources with Sections on Content Validity

Brown, J.D. & Hudson, T. (2002). *Criterion-referenced language testing.* Cambridge: Cambridge University Press.

Gorsuch, G. & Griffee, D.T. (2018). *Second language testing for student evaluation and classroom research.* Charlotte, NC: Information Age Publishing, Inc.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Feedback

Black, P., & Wiliam, D. (2012). Developing a theory of formative assessment. In: J. Gardner (Ed.), *Assessment and learning* (2nd ed.)(pp. 206-229). London: Sage Publications.

Fulcher, G. (2016). *Assessment for learning III: Effective feedback.* Retrieved: http://languagetesting.info/features/afl/formative3.html

Jang, E.E. & Wagner, M. (2014). Diagnostic feedback in the classroom. In J.A. Kunnan (Ed.), *Companion to language assessment* (693-711). John Wiley and Sons, Inc.

Sweetland Center for Writing (2019). *Providing feedback and grades to second language students.* Retrieved from: https://lsa.umich.edu/content/dam/sweetland-assets/sweetland-documents/teachingresources/ProvidingFeedbackandGradestoSecondLanguageStudents/ProvidingFeedbackAndGradesToSecondLanguageStudents.pdf

Wang, X. (2017). A Chinese EFL teacher's classroom assessment practices. *Language Assessment Quarterly, 14*(4), 312-327.

Formative Assessment/Classroom Assessment/Tests for Formative Purposes

Black, P. & Wiliam, D. (2012). Assessment for learning in the classroom. In J. Gardner (Ed.), *Assessment and learning* (2nd ed.) (pp. 11-32). London: Sage Publications.

Colby-Kelly, C. & Turner, C. (2007). AFL [Assessment for Learning] research in the L2 classroom and evidence of usefulness: Taking formative assessment to the next level. *The Canadian Modern language Review, 64*(1), 9-37.

Gardner, J. (2012). Assessment and learning: Introduction. In J. Gardner (Ed.), *Assessment and learning* (2nd ed.) (pp. 1-8). London: Sage Publications.

Tuttle, H.G. (2012). Daily formative assessments in second language acquisition. *Educator's Voice, V,* 58-63. Retrieved from: https://www.nysut.org/~/media/files/nysut/resources/2012/may/educators-voice-5-assessments/edvoicev_07_daily_assessment_esl.pdf?la=en

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Test Effect

Brame, C., & Biel, R. (2017). Test-enhanced learning: The potential for testing to promote greater learning in undergraduate science courses. *Life Sciences Education, 14*(2). Retrieved from: https://www.lifescied.org/doi/10.1187/cbe.14-11-0208

Roediger, H.L. & Karpike, J.D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*, 249-255.

Pedagogically Worthwhile Tests

Fulcher, G. (2010). *Open book-open web tests.* Retrieved from: http://languagetesting.info/features/open/book.html

Shepard, L. (2000a). The role of assessment in teaching and learning. *CSE Technical Report 517.* Los Angeles, CA: University of California, Los Angeles, Center for the Study of Evaluation.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

REFERENCES

Adair-Hauck, B., Glisan, E., & Troyan, F. (2013). *Implementing integrated performance assessment*. Alexandria, VA: ACTFL.

Alderson, J.C. (2006). The CEFR and the need for more research. *The Modern Language Journal, 91*, 659-663.

Alderson, J.C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Alderson, J.C., Figueras, N., Kuijper, H., & Nold, G., Takala, S., & Tardieu, C. (2006). Analysing tests of reading and listening in relation to the Common European Framework of Reference: The experience of the Dutch CEFR Construct Project. *Language Assessment Quarterly, 3*(1), 3-30.

American Council on the Teaching of Foreign Languages (2012a). *ACTFL proficiency guidelines 2012*. Retrieved October 24, 2018 from: https://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012

American Council on the Teaching of Foreign Languages (2012b). *Performance descriptors for language learners*. Retrieved October 30, 2018 from: https://www.actfl.org/publications/guidelines-and-manuals/actfl-performance-descriptors-language-learners

American Council on the Teaching of Foreign Languages (2018). Proficiency and assessment workshops. Alexandria, VA: ACTFL. Available: https://www.actfl.org/assessment-professional-development/professional-development-workshops/proficiency-assessment-workshops

American Council on the Teaching of Foreign Languages & National Council of State Supervisors for Languages (2018a). *Can-Do statements introduction*. Retrieved December 10, 2018 from: https://www.actfl.org/publications/guidelines-and-manuals/ncssfl-actfl-can-do-statements

American Council on the Teaching of Foreign Languages & National Council of State Supervisors for Languages (2018b). *NCSSFL-ACTFL Can-Do statements proficiency benchmarks*. Retrieved December 10, 2018 from: https://www.actfl.org/sites/default/files/CanDos/Can-Do_Benchmarks_Indicators.pdf

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Bachman, L. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L. (2002). Some reflections on task-based language performance assessment. *Language Testing, 19*(4), 453-476.

Bachman, L. & Palmer, A. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bachman, L. & Palmer, A. (2010). *Language assessment in practice.* Oxford: Oxford University Press.

Bachman, L. & Savignon, S. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL Oral Interview. *The Modern Language Journal, 70*(4), 380-390.

Balogh, S. & Lindsay, J. (Eds.) (2017). *Mind the culture gap (*2nd ed.). Tokushima, Japan: Care for Tokushima, Tokushima ken kyoiku insatsu kabushiki kai (Tokushima Prefectural International Conference).

Bandura, A. (1997). *Self-efficacy: The exercise of control.* New York: W.H. Freeman.

Barrette, C. (2004). An analysis of foreign language achievement test drafts. *Foreign Language Annals, 37*(1), 58-69.

Baztán, A. M., Torrecillas, A.B., Cuadrado, A.S., Guerrero, M.D. C., & Molero, S. S. (2015). Descripción y análisis del examen online de acreditación de dominio de español de los niveles B1 y B2: el eLADE. *CertiuniJournal, (1),* 52-66.

Beare, K. (2019). *Teaching to the test in ESL class.* Retrieved from: https://www.thoughtco.com/teaching-to-the-test-in-esl-class-4116840?print

Benigno, V. & do Jong, J. (2016). The "global scale of English learning objectives for young learners": A CEFR-based inventory of descriptors. In M. Nikolov (Ed.), *Assessing young learners of English: Global and local perspectives* (pp. 43-64). Cham, Switzerland: Spring International Publishing.

Beretta, A. (1991). Theory construction in SLA: Complementarity and opposition. *Studies in Second Language Acquisition, 13*, 493-511.

Bloom, B.S. (1976). *Human characteristics and school learning.* New York: McGraw Hill.

Boblett, N. (2012). Scaffolding: Defining the metaphor. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics, 12*, 1–16.

Borg, S. (1999). Teachers' theories in grammar teaching. *ELT Journal, 53*(3), 157-167.

Brasil Ministério da Educação, Secretaria da Educação Básica [Brazil – Ministry of Education, Basic Education Secretary]. (2016). *Base*

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

nacional comum curricular [National Common Curricular Base]. Brasília, DF.

Brindley, G. (1997). Assessment and the language teacher: Trends and transitions. *The Language Teacher Online, 21*(9). Retrieved May 11, 2018: http://jalt-publications.org/old_tlt/files/97/sep/brindley.html

Brown, J.D. (2005). *Testing in language programs.* New York: McGraw Hill.

Brown, J.D. & Hudson, T. (2002). *Criterion-referenced testing.* Cambridge: Cambridge University Press.

Brumen, M., Zupancic, T., Aguero, M.F., & Alonso-Belmonte, I. (2018). Foreign language teachers' feedback practices: A comparative study. *The New Educational Review, 53*(3), 216-226.

Bullough, R. V., Jr. (1989). First-year teacher: A case study. New York: Teachers College Press.

Butler, A. & Roediger III, H. (2007). Testing improves long-term retention in a simulated classroom setting. *Journal of Cognitive Psychology, 4-5,* 514-527.

Canale, M. (1988). The measurement of communicative competence. *Annual Review of Applied Linguistics, 8,* 67-84.

Celce-Murcia, M., Dornyei, Z., & Thurrell, S. (1995). Communicative competence: A pedagogically motivated model with content specifications. *Issues in Applied Linguistics, 6*(2), 5-35.

Center for Open Educational Resources on Language Learning (2010). *Lesson 2: ACTFL guidelines and national standards.* University of Texas at Austin. Retrieved May 30, 2018: https://coerll.utexas.edu/methods/modules/teacher/02/

Cheng, L., Rogers, T., & Huiqin, H. (2004). ESL/EFL instructors' classroom assessment practices: Purposes, methods, and procedures. *Language Testing, 21*(3), 360-389.

Christison, M., & Murray, D.E. (2014). *What English language teachers need to know: Volume III – Designing curriculum.* New York: Routledge.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Thousand Oaks, CA: Sage.

Clarke, M.A. (1994). The dysfunctions of the theory/practice discourse. *TESOL Quarterly, 28*(1), 9-26.

Clementi, S. & Terrill, L. (2017). *The keys to planning for learning.* Alexandria, VA: ACTFL.

Colby-Kelly, C. & Turner, C. (2007). AFL [Assessment for Learning] research in the L2 classroom and evidence of usefulness: Taking

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

formative assessment to the next level. *The Canadian Modern Language Review, 64*(1), 9-37.

Cope, B., & Kalantzis, M. (2009). "Multiliteracies": New literacies, new learning. *Pedagogies: An international journal, 4*(3), 164-195.

Council of Europe (2001). *Common European Framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press. Retrieved October 24, 2018 from: https://rm.coe.int/1680459f97

Council of Europe (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*. Retrieved December 1, 2018 from: https://rm.coe.int/1680667a2d

Council of Europe (2018). *Common European Framework of Reference for Languages: Learning, teaching, assessment companion volume with new descriptors*. Retrieved December 25, 2018 from: https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

Council on International Educational Exchange (CIEE). (2019). *TOEFL ITP*. Retrieved from: https://www.cieej.or.jp/toefl/itp/about.html

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238.

D2L. (2019). Retrieved January 15, 2019 from https://d2l.arizona.edu

Davies, A. (1990). *Principles of language testing*. Oxford: Blackwell.

Davies, A., Brown, A., Elder, C., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing*. Cambridge: Cambridge University Press.

Davison, C. (2004). The contradictory culture of teacher-based assessment: ESL teacher assessment practices in Australian and Hong Kong secondary schools. *Language Testing, 21*(3), 305-334.

De Silva, R. (2014). Rubrics for assessment: Their effects on ESL students' authentic task performance. Center for English language communication 4th symposium proceedings. Singapore: National University of Singapore. Available: http://www.nus.edu.sg/celc/research/books/4th%20Symposium%20proceedings/

Doe, C. (2015). Student interpretations of diagnostic feedback. *Language Assessment Quarterly, 21*(1), 110-135.

Donato, R. & Adair-Hauck, B. (2002). The PACE model: A story-based approach to meaning and form for standards-based language learning. *The French Review*. 76, 265-296.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Downing, S.M. (2006). Twelve steps for effective test development. In S.M. Downing & T.M. Haladyna (Eds.), *Handbook of test development* (pp. 3-25). Mahwah, NJ: Lawrence Erlbaum Associates.

Edelenbos, P. & Kubanek-German, A. (2004). Teacher assessment: The concept of 'diagnostic competence.' *Language Testing, 21*(3), 259-283.

Educational Testing Service (2019a). *TOEFL*. Author: Retrieved January 29, 2019 from: https://www.ets.org/toefl

Educational Testing Service (2019b). *TOEFL Junior*: *Test content*. Author. Retrieved January 29, 2019 from: https://www.ets.org/toefl_junior/content

Eportal.editoria.3. (2016). La historia de "La Pola." Colombia Aprende. Retrieved from: http://aprende.colombiaaprende.edu.co/es/agenda/efem%C3%A9rides/la-historia-de-%E2%80%98la-pola%E2%80%99

Figueras, N. (2012). The impact of the CEFR. *ELT Journal, 66*(4), 477-485.

Fox, C. (1993). Communicative competence and beliefs about language among graduate teaching assistants in French. *The Modern Language Journal, 77*(3), 313-324.

Frain, T. (2009). *A comparative study of Korean university students before and after a criterion referenced test* (Unpublished master's thesis). University of Southern Queensland, Australia.

Frodesen, J., & Eyring, J. (2007). *Grammar dimensions 4: Form, meaning, and use* (4th ed.). Boston, MA: Heinle Cengage Learning.

Fulcher, G. (1996). Invalidating validity claims for the ACTFL oral rating scale. *System, 24*(2), 163-172.

Fulcher, G. (1998). Widdowson's model of communicative competence and the testing of reading: An exploratory study. *System, 26*, 281-302.

Fulcher, G. (2004). Deluded by artifices? The Common European Framework and harmonization. *Language Assessment Quarterly, 1*(4), 253-266.

Fulcher, G. (2010a). *Practical language testing*. London: Hodder Education.

Fulcher, G. (2010b). Assessment for learning I: An introduction. Retrieved from: http://languagetesting.info/features/afl/formative.html

Fulcher, G. & Davidson, F. (2007). *Language testing and assessment*. London: Routledge.

García Glick, B. (2016). Improvement of present subjunctive oral production in graded VoiceThread tasks. *Journal of Teacher Action Research, 3*(1), 91-106.

Goh, C. & Aryoust, S.V. (2010). Investigating the construct validity of the MELAB Listening Test through the Rasch Analysis and Correlated

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Uniqueness Modeling. In J.S. Johnson, E. Lagergren, & I. Plough (Eds.). *Spaan Fellow working papers in second ore foreign language assessment* (pp. 31-68). Ann Arbor, MI: English Language Institute, University of Michigan.

Golombek, P. (1998). A study of language teachers' personal practical knowledge. *TESOL Quarterly, 32*(3), 447-464.

Goodier, T. (2014). *Working with CEFR* can-do *statements: An investigation of UK English language teacher beliefs and published materials*. British Council ELT Master's Dissertation Awards Winner. Retrieved December 30, 2018 from: https://englishagenda.britishcouncil.org/sites/default/files/filefield_paths/working_with_cefr_can-do_statements_v2_1.pdf

Google Drive (2019). Retrieved January 15, 2019 from https://drive.google.com

Gorsuch, G.J. (2007). Developing 'the course' for college level English as a foreign language learners and faculty members in Vietnam. *Asian EFL Journal Quarterly, 9*(1), 195-226.

Gorsuch, G. & Griffee, D.T. (2018). *Second language testing for student evaluation and classroom research.* Charlotte, NC: Information Age Publishing.

Graham, S. & Weiner, B. (1995). Theories and principles of motivation. In D. Berliner & R. Calfee (Eds.). *Handbook of educational psychology* (pp. 63-84). New York: MacMillan Library Reference USA.

Grice, H. P. (1975). Logic and conversation. In P. Cole, P. & J. L. Morgan (Eds.) *Syntax and semantics volume 3 speech acts* (pp. 41-58). New York: Academic Press.

Griffee, D.T. (2012a). The role of theory in TA and ITA research. In G. Gorsuch (Ed.), *Working theories for teaching assistant development* (pp. 39-61). Stillwater, OK: New Forums Press.

Griffee, D.T. (2012b). Using grounded theory to develop emergent explanations on how second and foreign language TAs construct their teacher theory. In G. Gorsuch (Ed.), *Working theories for teaching assistant development* (pp. 201-230). Stillwater, OK: New Forums Press.

Griffee, D.T. (2018). *An introduction to second language research methods: Design and data* (2nd ed.). Berkeley, CA and Kyoto, Japan: TESL-EJ Publications.

Griffee, D.T. & Gorsuch, G. (2016). *Evaluating second language courses.* Charlotte, NC: Information Age Publishing.

Griffee, D.T. & Gorsuch, G. (1999). The role of theory in ESL/EFL. *The Language Teacher, 23*(10), 35-37.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Hambleton, R. & Sireci, S. (1997). Future directions for norm-referenced and criterion-referenced achievement testing. *International Journal of Educational Research, 27*(5), 379-393.

Harsch, C. (2014). General language proficiency revisited: Current and future issues. *Language Assessment Quarterly, 11,* 152-169.

Hartig, J., Klieme, E., & Leutner, D. (2008). *Assessment of competencies in educational contexts.* Göttingen: Hogrefe.

Hatasa, Y. & Watanabe, T. (2017). Japanese as a second language assessment in Japan: Current issues and future directions. *Language Assessment Quarterly, 14*(3), 192-212.

Hawkes, M. (2012). Using task repetition to direct learner attention and focus on form. *English Language Teaching Journal, 66*(3), 327-336.

Herdman, B. (2017). The edamame experience. In S. Balogh & J. Lindsay (Eds.). *Mind the culture gap* (2nd ed.) (p. 59). Tokushima, Japan: Care for Tokushima, Tokushima ken kyoiku insatsu kabushiki kai (Tokushima Prefectural International Conference).

Hom, C., Ramos, M., Blumer, I. & Madaus, G. (2000). *Cut scores: Results may vary.* The National Board on Educational Testing and Public Policy. Retrieved June 17, 2019 from: https://www.researchgate.net/publication/28799837_Cut_Scores_Results_May_Vary_NBETPP_Monographs_Volume_1_Number_1

Hughes, A. (2010). *Testing for language teachers* (2nd Ed.). Cambridge: Cambridge University Press.

Hulstijn, J.H. (2011). Language proficiency in native and nonnative speakers: An agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly, 8*(3), 229–249.

Huntley, H. (2006). *Essential academic vocabulary: Mastering the complete academic word list.* New York: National Geographic Learning/Heinle Cengage Learning.

Hurt, N. (2019). *4 handy types of assessment in language teaching that stray from tradition.* Retrieved from: https://www.fluentu.com/blog/educator/assessment-in-language-teaching-4/#

Hymes, D. (1972). On communicative competence. In J.B. Pride & J. Holmes (Eds.), *Sociolinguistics: Selected readings* (pp. 269-293). Harmondsworth: Penguin.

IELTs (2018). *International English Language Testing System.* Author. Retrieved January 29, 2019 from: https://www.ielts.org/en-us

Instituto Cervantes (2019). *Exámenes DELE B1.* Retrieved from: https://examenes.cervantes.es/es/dele/examenes/b1

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Japan Association of College English Teachers (2016). *JACET 8000*. Retrieved February, 2019 from: https://www.jacet.org/publication/other-publications/

The Japan Times (2018, February 2). English proficiency at Japan's middle and high schools falls far short of government targets. Retrieved February, 2019 from: https://www.japantimes.co.jp/news/2018/04/07/national/social-issues/japans-high-school-english-proficiency-falls-short-government-targets/#.XEv48s8zb0Y

Jones, R.L. (1985). Second language performance testing: An overview. In P.C. Hauptman, R. LeBlanc, & M.B. Wesche (Eds.), *Second language performance testing* (pp. 15-24). Ottawa: University of Ottawa Press.

Jonsson, A. & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences, *Educational Research Review, 2*, 130–144.

Kato, K. (2006). English reading course in college: Translation method or extensive reading. *Bulletin of the Faculty of Literature of Aichi Prefectural University, 55*, 87-99.

Kern, R. (2000). *Literacy and language teaching*. Oxford: Oxford University Press.

Kern, R. (2008). Making connections through texts in language teaching. *Language Teaching, 41*(3), 367-387.

Kikuchi, K. (2005). Student and teacher perceptions of learning needs: A cross analysis. *Shiken: JALT Testing & Evaluation SIG Newsletter, 9*(2), 8-20.

Kissling, E. & O'Donnell, M. (2015). Increasing language awareness and self-efficacy of FL students using self-assessment and the ACTFL proficiency guidelines. *Language Awareness, 24*(4), 283-302.

Kramsch, C. (2006). From communicative competence to symbolic competence. *The Modern Language Journal 30*(2), S. 249–252.

Kramsch, C. (2009). *The multilingual subject*. New York: Oxford University Press.

Kramsch, C. (2011). The symbolic dimensions of the intercultural. *Language Teaching* 44(3), S. 354–367.

Krashen, S. (1982). *Principles and practice in second language acquisition*. New York: Pergamon.

Kunnan, A.J. (1998). Approaches to validation in language assessment. In A.J. Kunnan (Ed.), *Validation in language assessment* (pp. 1-16). Mahwah, NJ: Lawrence Erlbaum Associates.

Lantolf, J. & Frawley, W. (1988). Proficiency: Understanding the construct. *Studies in Second Language Acquisition, 10*(2), 181-195.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Leeming, F. (2002). The exam-a-day procedure improves performance in psychology classes. *Teaching of Psychology, 29*, 210-212.

Linacre, J. M. (2014). *A user's guide to WINSTEPS® MINISTEP*. Rasch-Model Computer Programs, Program Manual 3.81.0. Retrieved from: https://www.winsteps.com/manuals.htm

Liskin-Gasparro, J. (2003). The ACTFL Proficiency Guidelines and the oral proficiency interview: A brief history and analysis of their survival. *Foreign Language Annals, 36*(4), 483-490.

Livingston, S.& Zieky, M. (1982). *Passing scores: Standards of performance on educational and occupational tests.* Princeton, NJ: Educational Testing. Service.

Luoma, S. (2004). *Assessing speaking.* Cambridge: Cambridge University Press.

MLA Ad Hoc Committee on Foreign Languages (2007). *Foreign languages and higher education: New structures for a changed world.* Retrieved January 20th from: https://apps.mla.org/pdf/forlang_news_pdf.pdf

Manley, J. (1995). Assessing students' oral language: One school district's response. *Foreign Language Annals, 28*(1), 93-102.

Martinez-García, M. T. (2016). *Tracking bilingual activation in the processing and production of Spanish stress.* Doctoral dissertation, University of Kansas.

Maxim, H. (2002). A study into the feasibility and effects of reading extended authentic discourse in the beginning German language classroom. *Modern Language Journal, 86*(1), 20-35.

Maxim, H. (2006). Integrating textual thinking into the introductory college-level foreign language classroom. *Modern Language Journal, 90*(1), 19-32.

McNamara, T. (1997). Performance testing. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education: Volume 7 Language testing and assessment* (pp. 131-139).

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: MacMillan.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational measurement: Issues and practice, 14*(4), 5-8.

Moeller, J., Berger, S., Wieden, A., Mabee, B., & Adolph, W. R. (2017). *Kaleidoskop: Kultur, literatur und grammatik* (9th ed.). Boston: Cengage.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Moodie, I. & Haany, D. (2008). Using pair work exams for testing in ESL/EFL conversation classes. *The Internet TESL Journal, 14*(8), 1-6. Available: http://iteslj.org/Techniques/Moodie-PairWorkTesting.html

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation, 7*(10). Retrieved from: https://pareonline.net/getvn.asp?v=7&n=10

Murphey, T. (2001). Exploring conversational shadowing. *Language Teaching Research, 5*(2), 128-155. doi:10.1177/136216880100500203

National Foreign Language Research Center (NFLRC)(2019). *Fundamentals of Project-based Language Learning Online Institute.* Retrieved from: http://nflrc.hawaii.edu/projects/view/2014A/

National Standards Collaborative Board (2015). *World-readiness standards for learning languages.* Retrieved January 15 2019 from: https://www.actfl.org/sites/default/files/publications/standards/World-ReadinessStandardsforLearningLanguages.pdf

Newfields, T. (2006). Teacher development and assessment literacy. *Authentic communication: Proceedings of the 5th Annual JALT Pan-SIG Conference* (pp. 48-73). Shizuoka, Japan: Tokai University College of Marine Science.

New London Group (1996). A pedagogy of multiliteracies: Designing social futures. *Harvard Educational Review, 66*(1), 60-92.

Nikolov, M. (2016). A framework for young EFL learners' diagnostic assessment: 'Can do statements' and task types. In M. Nikolov (Ed.) Assessing young learners of English: Global and local perspectives (pp. 65-92). New York: Springer.

Norris, J.M. (2008). *Validity evaluation in language assessment.* Frankfurt, Germany: Peter Lang.

Norris, J.M., Brown, J.D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments.* Honolulu, HI: University of Hawaii Press.

North, B. (1995). The development of a common framework scale of descriptors of language proficiency based on a theory of measurement. *System, 23*(4), 445-465.

Norton, B. & Stein, P. (1998). Why the "Monkey's Passage" bombed: Tests, Genres, and Teaching. In A.J. Kunnan (Ed.). *Validation in language assessment* (pp. 231-249). Mahwah, NJ: Lawrence Erlbaum Associates.

Nummikoski, M. (2011). *Troika: A communicative approach to Russian language, life, and culture* (2nd ed.). Danvers, MA: John Wiley and Sons, Inc.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Nuttall, C. (1982). *Teaching reading skills in a foreign language.* London: Richard Clay, Ltd.

O'Sullivan, B. & Weir, C. (2011). Test development and validation. In B. O'Sullivan (Ed.), *Language testing: Theories and practices* (pp. 13-32). Houndsmills, England: Palgrave Macmillan.

Paesani, K., Allen, H. W., & Dupuy, B. (2015). *A multiliteracies framework for collegiate foreign language teaching.* Upper Saddle River: Pearson.

Patton, M.D. (2011). Mapping the gaps in services for L2 writers. *Across the disciplines, 8*(4). Available: http://wac.colostate.edu/atd/ell/patton.cfm

Pearson Education (2019). *My Spanish Lab.* Retrieved: https://www.pearsonmylabandmastering.com/northamerica/mylanguagelabs/

Phillips, J. & Abbott, M. (2011). *A decade of foreign language standards.* Grant Report #P017A080037 U.S. Department of Education. Alexandria, VA: American Council on the Teaching of Foreign Languages.

Piepho, H.E. (1974). *Kommunikative kompetenz als übergeordnetes lernziel im Englischunterricht.* Domburg/Frickhofen: Frankonius.

Piepho, H.E. (1979). *Kommunikative didaktik des Englischunterrichts.* Dornburg/Frickhofen: Frankonius.

Pratt, M., Geisler, M., Kramsch, C., McGinnis, S., Patrikis, P., Ryding, K., & Saussy, H., (2008). Transforming college and university foreign language departments. *The Modern Language Journal, 92*(2), 287-292.

Purpura, J. (2016). Second and foreign language assessment. *The Modern Language Journal, 100* (Supplement 2016), 190-208.

Rea-Dickins, P. (2004). Understanding teachers as agents of assessment. *Language Testing, 21*(3), 249-258.

Ringvald, V. (2006) Proficiency-based instruction. *Journal of Jewish Education, 69,* 5-8.

Roediger III, H., Agarwal, P., McDaniel, M., & McDermott, K. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology, 17*(4), 382-395.

Roediger III, H., & Karpicke, J. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17,* 249-255.

Ronald, J., Rinnert, C., Boyd, K., & Knight, T. (Eds.). (2012). *Pragtivities: Bringing pragmatics to second language classrooms.* Tokyo: Japan Association for Language Teaching, Pragmatics Special Interest Group.

Rost, M. (2002). *Teaching and researching listening.* London: Longman.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

Runnels, J. (2014). An exploratory reliability and content analysis of the CEFR-Japan's A-level Can-do statements. *JALT Journal, 36*(1), 69-89.

Savignon, S. (1985). The evaluation of communicative competence: The ACTFL provisional proficiency guidelines. *The Modern Language Journal, 69*(2), 129-134.

Schechtman, R. & Koser, J. (2008). Foreign languages and higher education: A pragmatic approach to change. *The Modern Language Journal, 92*(2), 309-312.

Schmitt, N., & Schmitt, D. (1995). Eleven principles for designing a vocabulary curriculum. *English Language Teaching Journal.*

Shehadeh, A. (2002). Comprehensible output, from occurrence to acquisition: An agenda for acquisitional research. *Language Learning, 52*(3), 597-647.

Shepard, L. (2005, November). Linking formative assessment to scaffolding. *Educational Leadership,* 66-70.

Shepard, L. (2000a). The role of assessment in teaching and learning. *CSE Technical Report 517.* Los Angeles, CA: University of California, Los Angeles, Center for the Study of Evaluation.

Shepard, L. (2000b). The role of assessment in a learning culture. *Educational Researchers, 29*(7), 4-14.

Siegle, D. (2000). *An introduction to self-efficacy.* Retrieved October 24, 2018 from https://nrcgt.uconn.edu/underachievement_study/self-efficacy/se_section1/

Snow, R., Como, L., & Jackson, D. (1995). Individual differences in affective and conative functions. In D. Berliner & R. Calfee (Eds.). *Handbook of educational psychology* (pp. 243-310). New York: MacMillan Library Reference USA.

Swaffar, J. & Arens, K. (2005). *Remapping the foreign language curriculum: an approach through multiple literacies.* New York: The Modern Language Association of America.

Taguchi, E. & Gorsuch, G. (2002). Transfer effects of repeated EFL reading on reading new passages: A preliminary investigation. *Reading in a Foreign Language, 14*(1), 43-65.

Taipei Economic and Cultural Office (2011). *TOCFL.* Los Angeles, CA: Author. Retrieved on January 29, 2019 from: http://www.tw.org/tocfl/

University of Cambridge Local Examinations Syndicate (UCLES) (2019). *Cambridge Assessment English.* Retrieved from: https://www.cambridgeenglish.org

Van den Branden, K. (1997). Effects of negotiation on language learners' output. *Language Learning, 47*(4), 589-636.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

van Lier, L. (1996). *Interaction in the language curriculum.* New York: Longman.

van Lier, L. (2004). *The ecology and semiotics of language learning: A sociocultural perspective.* Boston: Kluwer Academic Publishers.

Venema, J. (2002). Developing classroom specific rating scales: Clarifying teacher assessment of oral communicative competence. *Shiken: JALT testing & evaluation SIG newsletter, 6*(1), 2-6.

VoiceThread (2019). *VoiceThread.* Retrieved from: https://voicethread.com/

Wang, X. (2017). A Chinese EFL teacher's classroom assessment practices. *Language Assessment Quarterly, 14*(4), 312-327.

Wang, Y. (2008). Influence on planning on students' language performance in task-based teaching. *English Language Teaching, 1*(1), 83-86. Retrieved from: https://files.eric.ed.gov/fulltext/EJ1082592.pdf

Warner, C. (2011). Rethinking the role of language study in internationalizing higher education. *L2 Journal, 3*(1), 1-21.

Widdowson, H. (1983). *Learning purpose and language use.* Oxford: Oxford University Press.

Weinert F.E. (2001). Vergleichende Leistungsmessung in Schulen - eine umstrittene Selbstverständlichkeit. In Weinert F.E. (Hg.): *Leistungsmessung in Schulen.* Weinheim/Basel: Beltz S. 17-33.

Winchester, S. (2004). *Krakatoa.* New York: Penguin Books.

Xreading (2019). Retrieved from: https://xreading.com/

Zieky, M. & Perie, M. (2004). *A primer on setting cut scores on tests of educational achievement.* Princeton, NJ: Educational Testing Service. Retrieved June 17, 2019 from: https://www.ets.org/Media/Research/pdf/Cut_Scores_Primer.pdf

Zimmerman, B. & Schunk, D. (2008). Motivation: An essential dimension of self-regulated learning. In D. Schunk & B. Zimmerman (Eds.). Motivation and self-regulated learning (pp. 1-30). New York: Lawrence Erlbaum Associates.

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

INDEX

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use

EBSCOhost - printed on 2/10/2023 11:39 AM via . All use subject to https://www.ebsco.com/terms-of-use