Language Assessment Literacy

From Theory to Practice

Edited by

Dina Tsagari

Language Assessment Literacy

Language Assessment Literacy:

From Theory to Practice

Edited by

Dina Tsagari

Cambridge Scholars Publishing

Language Assessment Literacy: From Theory to Practice

Edited by Dina Tsagari

This book first published 2020

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

ISBN (10): 1-5275-4835-X
ISBN (13): 978-1-5275-4835-0

TABLE OF CONTENTS

INTRODUCTION

The field of language testing and assessment (LTA) has recently recognized the importance and underlying theoretical and practical underpinnings of Language Assessment Literacy–LAL, an area that is gradually coming into its own with an evolving research agenda (Taylor 2013; Pill and Harding 2013; Inbar-Lourie 2008; Fulcher 2012, Scarino 2013; Vogt and Tsagari 2014).

This book is an edited volume based on the collection of theoretical and research papers within LAL titled "Language Assessment Literacy: Theory and Practice". The book was motivated by the international conference organized as part of the dissemination of an EU-funded project that aimed at promoting teachers' LAL: "Teachers' Assessment Literacy Enhancement – TALE" Erasmus+, Cooperation for innovation and the exchange of good practices, Strategic Partnerships for school education, KA2, European Lifelong Learning Programmes, 2015-1-CY01-KA201-011863 (http://taleproject.eu/). The chapters in the volume comprise papers that were presented during the conference and others that were recruited through invitations sent out via online LTA networks and professional lists.

The volume addresses the notion of LAL through a multidisciplinary and multifaceted approach for the wide language assessment research, teaching and academic community and beyond. The chapters of the volume examine various LAL issues that range from definitions of assessment literacy, conceptualizations and research in LAL to relationships between LAL and teachers' assessment practices, test washback, assessment of young learners and gifted students, computer-assisted language assessment, alternative assessment paradigms, and specific language skills such as reading and oracy skills.

The book comprises 14 chapters organized in three thematic areas:

Part I: Defining Language Assessment Literacy (2 chapters)
Part II: Understanding Language Assessment Literacy (5 chapters)
Part III: Expanding Language Assessment Literacy (7 chapters)

The chapters intertwine and entangle these issues from different perspectives. Contributors to the volume address the issues identified from theoretical and empirical points of view with important implications for teachers as the critical stakeholder group. Studies of both a cross-sectional and longitudinal nature are included as well as studies conducted with young and adult test takers in either high- or low-stakes environments. The languages examined are many and varied, such as English, German, French, Italian, Spanish, Russian, and Greek, and so are the contexts of inquiry: Austria, Cyprus, Greece, Norway, Tunisia, and Russia.

The first two chapters outline and reflect on the concept of language assessment literacy. Specifically, in the first chapter, Glenn Fulcher builds upon his previous research (Fulcher 2012) to extend the concept of LAL and presents a model of LAL acquisition for language teachers. The model has a philosophical basis and is illustrated with practical examples of materials and activities within the context of an MA programme for language teachers. In the second chapter, Gudrun Erickson touches on a number of very important issues that can be used for and by teachers (pre- and in-service), students and other stakeholders and provides concrete examples of international sources and resources for the enhancement of teachers' language assessment literacy.

Underpinned by considerations of a broad view of validity, the next three chapters focus on teachers' conceptions of assessment and assessment literacy in second language assessment. In the first of these chapters Anastasia Drackert, Carmen Konzett-Firth, Wolfgang Stadler and Judith Visser present selected results of an empirical study investigating language assessment practices and training of in-service secondary school teachers of French, Italian, and Spanish in Austria and Germany. The survey results present teachers' reported practices and their subjective theories on the purposes of assessment. The authors emphasize the need to involve teachers' subjective theories in assessment practices. In the following chapter Georgia Solomonidou, also stresses the same point that perceptions of assessment are important on building healthy systems of LAL in student assessment of the Modern Greek language as a first language. Through her sequential mixed-methods research she shows that the conceptions of secondary school teachers are in alignment with the current shift of assessment in the local context to be used for enhancing teaching and learning.

Christine-Nicole Giannikas in her chapter tells us that it is also important to consider the assessment practices of teachers of YLs. Data were gathered via a global online survey regarding the use of games as an assessment tool, teachers' perspectives, and assessment training. The outcomes show the need for the enhancement of LAL, and the need to include alternative assessment in particular in teacher training programmes.

The next chapter written by Olga Kvasova focuses on three LTA courses developed by the author and delivered in a classical university in Ukraine at undergraduate and graduate levels. The author discusses the course components and the methodology of course delivery, e.g. modes and formats of instruction and techniques of training and assessing the LTA knowledge and skills of participants. The findings of the study are of use to educators and institutions who aspire to enhance teachers' LAL at national and international levels.

This section of the book closes with a chapter by Yasemina Karagiorgi and Alexandra Petridou which presents a nationwide study that focuses on the Programme for Functional Literacy (PfL), implemented in Cyprus that aims to identify students "at risk" at Years 3 and 6 across all public primary schools of the country. The chapter argues in favour of maintaining in place the longitudinal national assessment programme as it enhances aspects of teachers' language assessment literacy.

The third part of the book comprises equally interesting chapters that research various aspects of LAL. For example, the first chapter of this section written by Asma Maaoui and Dina Tsagari examines teacher- and context-related factors in reading assessment in Tunisian higher education within the interface of LAL and test validity. The results show how intrinsic and extrinsic factors shape the teachers' LAL at the individual level and identify the challenges caused by inadequate teacher assessment literacy underlying the test design process.

In the north of Europe, Norway, Anne-Grete Kaldahl examines the lower secondary students' perceptions of the assessment of oracy across subjects, drawing on a rhetorical topos analysis of qualitative data generated through focus group interviews in two public schools. The results have important implications for teachers and their LAL enhancement.

Wolfgang Stadler and Anna Dreher look at sociopragmatic assessment literacy in the Russian language classroom. The authors examine coursebooks and test collections that reveal implications for the various stakeholders and offer useful information for improving teaching and assessment competence thus supporting the assessment literacy of key stakeholders.

Svetlana Karpava looks at how alternative assessment can be aligned with active, lifelong learning. In her study set in Cyprus the author investigates assessment in language courses for 20 university EFL tutors and 64 secondary school English teachers. The analysis of the data shows that the tutors like the idea of communities of shared practice and developing into reflective practitioners but factors such as restrictions of time and curricula aims are not conducive to raising levels of LAL.

Marina Perevertkina researches an equally important aspect of LAL, that of designing a conducive assessment environment for gifted students. The study, set in Herzen State Pedagogical University of Russia, reviews the content of a course that was specifically designed with this aim in mind and outlines the strategies in choosing valid techniques of assessment, in terms of developing, scoring and interpreting results and assessment methods appropriate for gifted students.

Irini-Renika Papakammenou sheds light on another very important aspect of language assessment, that of test washback and its relation with teachers' assessment literacy. The chapter suggests how to decrease or even eliminate negative washback and provides ways to employ communicative methodologies and alternative assessment techniques based on a research study in a multi-exam context in Greece.

The final chapter written by Jack Burston, Androulla Athanasiou and Maro Neophytou addresses the challenge of aligning vocabulary and grammar difficulty with CEFR proficiency ratings. The authors propose a computer-adaptive test of English, namely the E-CAT, which can considerably enhance LAL levels for its users.

In conclusion and given its scope and nature, this edited volume is of use to teachers and institutions committed to making assessment literacy sustainable in diverse national and international educational contexts. The book also hopes to chart new avenues of research in the field of LAL in the future and become an important reference source and reading material in undergraduate and postgraduate courses in teacher education in colleges and universities around the globe.

References

Fulcher, Glen. 2012. "Assessment Literacy for the Language Classroom." *Language Assessment Quarterly* 9(2):113-132.

Inbar-Lourie, Ofra. 2008. "Constructing a Language Assessment Knowledge Base: A Focus on Language Assessment Courses." *Language Testing* 25(3):385-402.

Pill, John, and Luke Harding. 2013. "Defining the Language Assessment Literacy Gap: Evidence from a Parliamentary Inquiry." *Language Testing* 30(3):381-402.

Scarino, Angela. 2013. "Language Assessment Literacy as Self-awareness: Understanding the Role of Interpretation in Assessment and Teacher Learning." *Language Testing* 30(3):309-327.

Taylor, Lynda. 2013. "Communicating the Theory, Practice and Principles of Language Testing to Test Stakeholders: Some Reflections." *Language Testing* 30(3):403-412.

Vogt, Karin, and Dina Tsagari. 2014. "Assessment Literacy of Foreign Language Teachers: Findings of a European Study." *Language Assessment Quarterly* 11(4):374-402.

PART I:

DEFINING LANGUAGE ASSESSMENT LITERACY

CHAPTER ONE

OPERATIONALIZING LANGUAGE ASSESSMENT LITERACY

GLENN FULCHER

Defining Language Assessment Literacy (LAL) is important; but only as a precursor to achieving it (Stiggins 1991). It is arguably the case that existing texts designed for use in language testing and assessment programmes attempt to do just this (Davies 2008), and one recent introductory text in assessment is subtitled "theory and practice in assessment literacy" (Manning 2016). Yet, there is little explicit articulation between survey research into LAL needs, definitions of the construct, and teaching/learning strategies, even in research that is explicitly aimed at developing learning programmes and materials (Huai et al. 2006; Berry and Sheehan 2017; Sheehan and Monro 2017). Furthermore, when it is claimed that attempts to develop LAL are based upon a named philosophy such as "constructivism", there is no attempt to demonstrate how the philosophy informs practice (see Conole et al. 2004, 17). This chapter builds upon the survey research of Fulcher (2012) to extend the definition of LAL presented there into an overt theory of pedagogy that can inform teaching and learning for LAL in one specific context, namely the professional development of practising language teachers undertaking the language testing modules on an MA programme. The approach described is embedded within a pragmatic theory of learning through doing that draws upon the metaphor of the apprentice. The resulting pedagogy is illustrated in practice through examples drawn from print and online learning materials. The chapter adds to the LAL debate by extending the discussion from definition and needs analysis to learning and teaching.

1. LAL Research

Traditional approaches to defining Assessment Literacy (AL) depend on classification rather than empirical research and tend to draw heavily upon psychometric theory and technical questions of validity (Bracey 2000; Popham 2009). This is encouraged by rather narrow definitions, such as: "Assessment literacy consists of an individual's *understandings* of the *fundamental assessment concepts* and *procedures* deemed *likely to* *influence educational decisions*" (italics in the original) (Popham 2011, 267). Similarly, a-priori "standards" have been produced upon which tests of AL have been created (Brookhart 2011). When these definitions or standards have been used as a basis for learning and teaching, they have led to very restricted course content that has changed little since Shaffer's (1993) survey of curriculum content for AL.

Research into LAL has largely avoided such a-priori definitions. It does, however, begin with the work of Stiggins (1991). Although he attempted to define what was meant by a person who is "assessment literate", he also raised the question of how assessment literate various stakeholders might be. These two questions have dominated LAL research to date, explaining the concentration of effort into questionnaire and survey design with the dual purpose of discovering the needs of specific groups of learners, and extracting from these needs a working definition of LAL (see Inbar-Laurie 2016).

Hasselgreen, Carlsen and Helness (2004) and Huhta, Hirvalä and Banerjee (2005) conducted a survey designed to uncover the assessment training needs of teachers in Europe, and reported that the most important areas for developing LAL are in the areas of portfolio assessment, classroom tests, peer- and self-assessment, feedback (all arguably assessment for learning), interpreting scores, validity, reliability, statistics, item writing, and rating performance tests (all more related to summative assessment). DeLuca and Klinger (2010) surveyed 700 trainee teachers in Ontario using a questionnaire based upon an a-priori understanding of LAL, arriving at the following definition based on factor analysis and labelling: "…the understanding and appropriate use of assessment practices along with the knowledge of the theoretical and philosophical underpinnings in the measurement of students' learning" (DeLuca and Klinger 2010, 429-420). Crusan et al. (2016) conducted an international survey of writing teachers, but this focused primarily on attitudes rather than needs and definitions. Sheehan and Munro (2017) conducted qualitative research into teacher attitudes to assessment, and their assessment needs, using a small convenience sample. However, no clear definition of LAL emerges. While there is a claim that the research was conducted as a preliminary to producing training materials there is no link between LAL theory and materials, and no discussion of pedagogy or principles to guide the construction of learning materials.

Fulcher (2012) conducted an online survey to discover the needs of language teachers, also conducting factor analysis on the closed items and expanding the factor definitions with reference to the qualitative accounts provided by individuals. He arrived at an "expanded definition" of LAL as:

> The knowledge, skills and abilities required to design, develop, maintain or evaluate, large-scale standardized and/or classroom-based tests, familiarity with test processes, and awareness of principles and concepts that guide and underpin practice, including ethics and codes of practice. The ability to place knowledge, skills, processes, principles and concepts within wider historical, social, political and philosophical frameworks in order to understand why practices have arisen as they have, and to evaluate the role and impact of testing on society, institutions, and individuals. (Fulcher 2012, 125)

While the generic definition of LAL provided by Fulcher (2012) seems to have been widely accepted as a useful starting point for future research (Tsagari and Vogt 2017, 41), the question of stakeholder literacy still preoccupies researchers. Taylor (2013) quite reasonably suggests that different profiles of LAL might be needed for different stakeholder communities, such as professional testers, teachers, and admissions tutors. As a result of Taylor's theoretical suggestions, researchers have concentrated on the development of ever more complex survey instruments in an attempt to provide an empirical basis to claims about the assessment literacy needs of the different stakeholder groups (Harding and Kremmel 2016). This avenue of research will continue to bear fruit, particularly when it seeks to elucidate the needs of identifiable groups of stakeholders, such as policy makers (Pill and Harding 2013), trainee teachers (DeLuca and Klinger 2010; Xu and Brown 2016; Hill 2017), students (Smith et al. 2011), and university admissions tutors (Deygers and Malone 2019), especially if new instruments can solve the sampling and psychometric issues outlined by Fulcher (2012) and confirmed as problematic in other studies (e.g., Xu and Brown 2017, 149).

However, Stiggins raises a third question that has so far not been addressed: the role of teaching and learning to achieve assessment literacy. He did not specify what should be learned, or how it was to be learned; but he considered teaching and learning to be the ultimate goal to be served by creating theoretical definitions and conducting needs analyses for stakeholder groups.

> What will it take to create a society that is assessment literate? We must start with the right attitudes and then undertake a great deal of basic assessment training. (Stiggins 1991, 538)

Little has been done to address this question, although where the outcomes of LAL training have been investigated the evidence suggests that there are significant problems with pedagogy (Lam 2015). In what follows I present a model of LAL acquisition for language teachers as a critical stakeholder group. The language teachers are undertaking two language testing modules as part of their MA degree in Applied Linguistics and TESOL. The model has an explicit philosophical basis, illustrated with practical examples of materials and activities. Both the philosophical basis and the practical materials may not be the most appropriate (or even relevant) to other stakeholder groups, but they may nevertheless act as a starting point for considering other targeted pedagogies.

2. An Apprenticeship Model

I argue that literacy in language testing and assessment for language teachers is about designing and building tests for their own use, and the institutions for which they work. Many will also work with local or national educational authorities to build tests for wider use. It is therefore appropriate to use the metaphor of the apprentice for this audience.

While the modern conception of apprenticeship is purely practical, its original meaning included a grounding in theory and an understanding of society, as well as an ability to make things. This is echoed in Fulcher and Davidson (2007):

> The practice of language testing draws upon, and also contributes to, all disciplines within applied linguistics. However, there is something fundamentally different about language testing. Language testing is all about building better tests, researching how to build better tests and, in so doing, understanding better the things that we test.
>
> Sociolinguists do not create ‘sociolinguistic things’. Discourse analysts do not create discourses. Phonologists do not create spoken utterances. Language testing, in contrast, is about *doing*. It is about *creating tests.*
>
> (Fulcher and Davidson 2007, xix. Italics in the original)

In ancient Greece craftsmen were referred to as “demioergoi”, derived from the two words for *public* (demios) and *work* (ergon). Hephaestus, the god of craftsmen, was responsible not only for the creation of objects, but also of civilization itself. There was no separation between theory and practice, knowledge and doing. Each informed the other. It was only in classical times that the link was severed. Aristotle changed how we see the world by referring to a craftsman as a “cheirotechnon”, literally translated as a “hand worker”. From then on artisans were seen as makers of things divorced from the theory that informs the making. The result was a binary class system with the artisans in the inferior role. By Roman times the rift was institutionalized in the hierarchical structure of society. In his *Ten Books of Architecture* (circa 20 BC), Vitruvius said that “The several arts are composed of two things–craftsmanship and theory. Craftsmanship belongs only to those who are trained…in the work; theory is shared with all educated persons” (cited in Lester and Piore 2004, 98). Making and building things were no longer considered thinking activities.

This rupture was healed to some degree during the Enlightenment. Diderot argued that it was necessary to understand theory and make mistakes in order to become a master craftsman. The process of error resulted in greater knowledge and understanding of the theory that informs quality work. “Become an apprentice and produce bad results so as to be able to teach people how to produce good ones” (cited in Sennett 2009, 96). The concept of quality and the desire to achieve it inevitably require understanding, setting standards (in the sense of what counts as a “quality product”), and searching for better ways of achieving desired goals.

> *Knowledge must come through action; you can have no test which is not fanciful, save by trial.* (Sophocles 496 BC to 406 BC)

This principle is illustrated most vividly in movements like the Arts and Crafts movement, and the workshops of great craftsmen like Stradivarius.

The metaphor of apprenticeship has four critical implications for education. Firstly, the learner is at the centre of the process. Secondly, the learner learns by doing. Thirdly, there is a master craftsman who is on hand to guide and mentor the learner. Fourthly, in their apprenticeship Peirce argues that (1878/1958, 328): “They should be made to feel that they are doing real and important work which was to appear in the digests of science and for the accuracy of which they are responsible…."

The notion of “responsibility” is essential for any apprenticeship model of learning. By doing and creating, the tests (or test fragments) produced by learners should be evaluated in terms of their usefulness for the purpose for which they were designed. Both the design effort and the evaluation are therefore to be informed by theory. But the apprenticeship model also implies a task-based pedagogy, which requires a pool of tasks that leads the apprentices to literacy through action. The structure and oversight are provided by the master craftsman, the person who is responsible for guiding learning, designing learning tasks, and providing critical feedback on both the product and process of the test item/task design.

This pedagogic model may encompass "reflective practice" and "community activities" (Xu and Brown 2016, 158), but it goes beyond them by proposing a structured pedagogic model that consists of apprenticeship tasks and the content to be learned. We deal with each of these in the next two sections.

3. Characteristics of Apprenticeship Tasks

Coomey and Stephenson (2001) conducted a meta-study to extract from research the key elements of successful online learning, which can be used as a starting point for understanding how both text- and web-based multimedia content can be used in LAL learning. They discovered from the range of research they reviewed that the structure and design of learning materials are considered successful when they engage four positive behaviours.

The first is the generation of *dialogue* around the input material and tasks. Interaction is varied by constructing either convergent or divergent goals for the participants (Pica et al. 1993, 13). Convergent tasks encourage collaboration and the achievement of specified outcomes. In language test development these are advantageous to show learners that test design cannot be done in isolation. Divergent tasks that resemble "game playing" with an element of competition have the capacity to generate challenges and critiques of ideas and practices, and they are particularly useful for considering opposing views of ethical test use or the role of tests in society.

The second is *involvement*, defined as the extent to which learners become engaged by the material, such that it generates intrinsic motivation. This is achieved if tasks are sufficiently challenging. While involvement is difficult to assess, I would argue that it grows out of *responsibility*: the feeling that valuable work is being done that will benefit the students, institutions and communities that language testers serve. This requires learners to engage with the social functions of language testing.

Thirdly it is necessary to provide *support* to learners while engaging with tasks. Feedback from other apprentices and the master is critical to learning. It is not an afterthought but integrated into the structure and progression of learning tasks (Black and Wiliam 2009).

Fourth is the structuring of *control* over an activity such that beginners are provided with more guidance, while more advanced learners are given more freedom to engage with tasks as they wish. This recognizes that with greater knowledge comes the freedom to innovate; with innovation comes error; with error comes learning.

These four characteristics may be referred to as the "**DISC** qualities" (**D**ialogue, **I**nvolvement, **S**upport and **C**ontrol); but as useful as this research might be to help define how a "good task" may engage learners, it is not sufficient for task design purposes. We also require a classification of task elements that provides a blueprint for design; similar, indeed, to a task specification template in language testing. There are many to choose from. However, one of the most useful, having stood the test of time, is that provided by Candlin (1987), which I have adapted for the purpose of LAL operationalization in Table 1 below.

Input	Stimulus to generate features of DISC
Roles	The assignment of participant duties within the task
Settings	The context for which test use if required
Actions	What participants must do to achieve goals
Outcomes	The goals of the task
Objectives	What you expect participants to learn (learning outcomes)
Feedback	Evaluation of performance and outcomes to inform iterative learning and improvement

Table 1: Classification of Elements of LAL Tasks, adapted from Candlin (1987)

This structure will be exemplified in the discussion of sample tasks below.

4. The Content of Apprenticeship Tasks

Developing a content model for apprenticeship tasks was the purpose of Fulcher (2012), whose definition of LAL is provided above. That definition is drawn from three aspects of LAL that are presented in Figure 1.

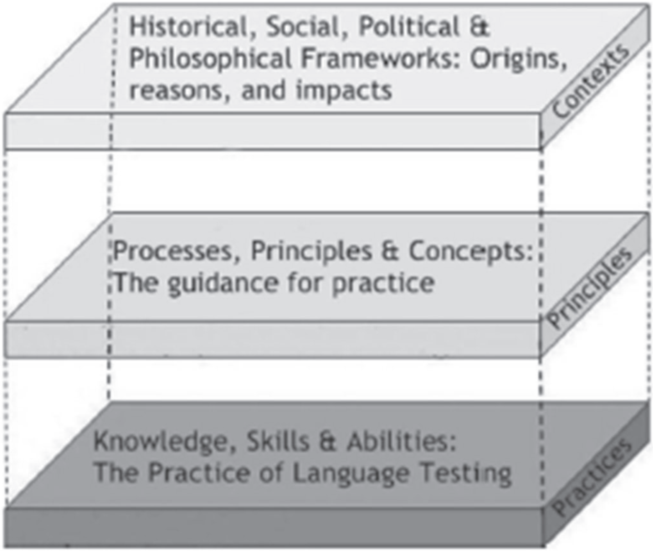


Figure 1: Language assessment literacy: An expanded definition (Fulcher 2012, 126)

Harding and Kremmel recognize the intention behind the apprenticeship content model:

> Importantly, Fulcher (2012) proposes that theoretical concepts in testing textbooks and courses should be presented within the context of practical test construction, using the test development cycle as the scaffold and introducing core principles and core terminological knowledge along the way rather than merely introducing them as decontextualized components of LAL. An implementation of this approach can be found in Fulcher's (2010) textbook *Practical Language Testing.* (Harding and Kremmel 2016, 419)

Additionally, Fulcher (2015) made a deliberate attempt to address the top level of the model presented in Figure 1. Essentially, the model became a plan for the content of a LAL programme realized through inter-related texts with supporting web-based materials.

The content model in Figure 1 is not intended to represent a hierarchy, but a movement from the foundation in practice to the abstraction of theory. The exploration of theory is realized in the practice of assessment design, using theory to inform practical decision-making. In practice, this is achieved by structuring tasks around the test design cycle as illustrated in Figure 2. Each element in the cycle is characterized by specific tasks, each of which involves decisions that are informed by theory and research.

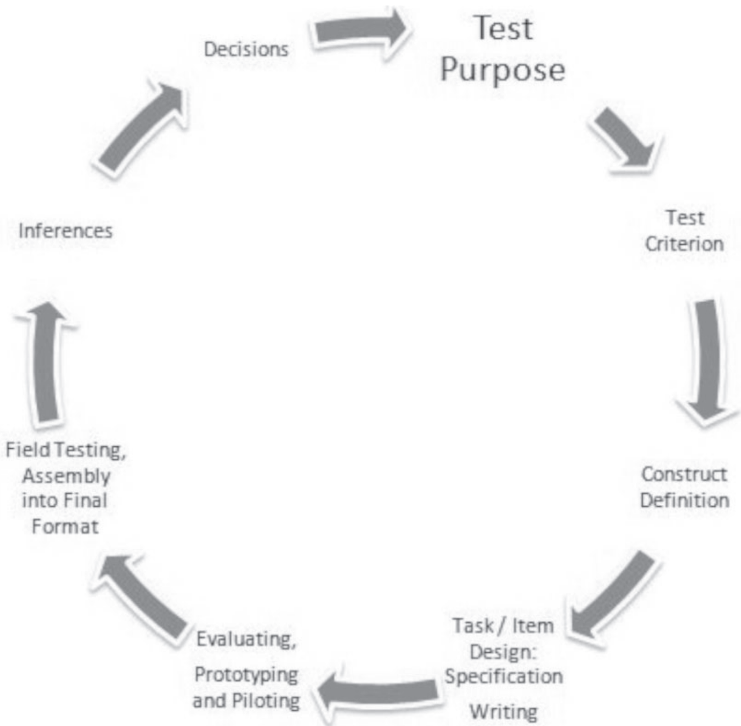


Figure 2: The Test Design Cycle (from Fulcher 2010, 94)

At each step in the cycle the aim is to create sets of learning tasks that implement DISC characteristics through tasks with explicit structures that aim to foster LAL in one or more elements of the content from the LAL model.

5. Illustrative Text-based Task Types

Below I present two tasks to illustrate the principles outlined above and relate each to both the characteristics and content of the apprenticeship model. The first would be used as part of the exploration of test purpose and use. The second is a design and review task that encourages the application of theory and data to item evaluation and revision. In each case, I make explicit the task elements from Table 1.

Task Type 1: **Debate** (Fulcher 2010, 23)

Content: Understanding social and political frameworks, impacts on individuals and society, through the exploration of a radical and controversial policy for a new use of educational assessments. Students have read Foucault, Mill, and a commentary on their differing views of tests before coming to the debate.

Sample Task

Read the following article. Do you believe that the creation of the database with individual dossiers for life is legitimate? Or does this just go to show that Foucault was right about the true intentions of governments? List the pros and cons on both sides of the argument. If you are working with a group of colleagues you may wish to organize a formal debate, with the motion "This house believes that a 'testing record for life' is an infringement of personal liberties and damaging to the future of the individual."

Adapted from *The Times*

Every child in school numbered for life

All 14-year-old children in England will have their personal details and exam results placed on an electronic database for life under a plan to be announced tomorrow.

Colleges and prospective employers will be able to access students' records online to check on their qualifications. Under the terms of the scheme all children will keep their individual number throughout their adult lives, The Times has learnt. The database will include details of exclusions and expulsions.

Officials said last night that the introduction of the unique learner number (ULN) was not a step towards a national identity card. But it will be seen as the latest step in the Government's broader efforts to computerise personal records.

Last night teachers' leaders, parents' organizations, opposition MPs and human rights campaigners questioned whether this Big Brother approach was necessary and said that it could compromise the personal security of millions of teenagers.

Stimulus: Media items and news reports of a controversial nature.
Roles: The proposer and seconder for the motion, the principal opponent. Other participants join in the debate on one side or the other.
Settings: Social and political issues of interest to a wider audience than just language testers. Issues of privacy, selection, privilege and individual rights are particularly suitable.
Actions: Group members consider the arguments for or against together; preparation of arguments and initial speeches outside the classroom; preparation of a presentation (transferable skills).
Outcomes: A debate lasting up to one hour, involving two groups of protagonists. Primarily involving divergent interactions.
Objectives: Deeper understanding of the role of test scores, qualifications, and records of other personal achievements in life chances and employment. Explore the ethics of personal data sharing and access.
Feedback: Post-debate debriefing, perhaps inviting all participants to "vote" on whether or not they would favour the introduction of a lifetime database.

Task Type 2: **Item Review** (Fulcher 2010, 192-194)

Content: Prototyping and evaluating new test items. This task type is situated within a series of tasks designed to sensitize students to potential problems with closed-response items before attempting to craft and prototype their own. The level of control is relatively high as the flaws in the items have been identified for their usefulness in representing common errors made by novice item writers.

Sample Task

The following four items were designed to test the pragmatic competence of intermediate level adolescent learners of English. Review each item and identify any flaws that you find. Each item may contain multiple problems.

Item 1

Tony: Finishing the packing for our holiday is going to take forever at this pace.
Linda: Yes, and we've had quite a few late nights recently.

Linda implies that they

(a) will miss their flight.
(b) will be up packing late into the night*.
(c) are both very tired.
(d) need to work faster.

Item 2

Presenter: So, what do people feel about binge drinking on the streets of our towns and cities, and particularly the rising incidence of drunkenness among young girls? We went out and about with our microphone to find out. Here's what Tom, an office worker from Middlington had to say.

Tom: Well, I mean, it's up to them isn't it? Okay, you know, perhaps they don't have the money, so they have to get it from somewhere, and it could damage their health. But it's what they want to do. So I don't see the problem.

Tom's view is that
(a) drinking causes social problems.
(b) young people need more money.
(c) heavy drinkers get liver disease.
(d) the young can do as they please*.

Item 3

Listen to the exchange and answer the question.

Rebecca: There are only three tickets for the concert on Friday, and I've invited Sonya and Terry.
Angela: I guess I'll get over it in time.

Angela
(a) is very upset.
(b) feels left out*.
(c) doesn't have time.
(d) has her own ticket.

Item 4

Economist: It is highly likely that the credit crunch will become excessively tighter as the year progresses, forcing more small- to mediumsized businesses into liquidation, and even resulting in many larger companies and high street brands being forced into the hands of the administrators. As governments become more involved with the banking sector many analysts foresee increasing levels of regulation that will bring an end to many opaque practices such as the trade in derivatives.

The economist argues that in the coming year
(a) we will all have a harder life.
(b) governments will buy banks.
(c) some businesses will close*.
(d) shops will get better managers.

Stimulus: Test items produced by colleagues or taken from existing tests. These need not be multiple choice.
Roles: All participants are critical evaluators. Ideally two or more groups should evaluate the items independently.
Settings: Prototyping new item types/evaluating existing items and tests.
Actions: Group members work convergently towards agreement on the strengths and weaknesses of each item.
Outcomes: An analysis of each item, and a decision about whether the item should be retained, revised, or rejected.
Objectives: Understanding what makes a "good" item. Acquiring the skills to engage in an item review, including key checks (for m/c items), a bias review, and an editorial review. Investigating the congruence of an item to the specification.
Feedback: Groups compare analyses, agree and/or critique the analyses of other groups.

This approach to operationalizing LAL has to date received favourable reviews (Read 2011, 304) because the application of the apprenticeship model creates a learner-centred learning environment in which theory informs decision-making by the learners. By doing so, they exercise judgement in order to arrive at responsible design decisions, taking into account the likely impact of their designs on individuals, institutions and society.

6. Illustrative Web-based Apprenticeship Tasks

Text-based tasks can target specific skills or knowledge in the test design process, but web-based tasks with their multi-media content can be used to structure integrated tasks that bring together many elements of the test-design cycle into a single activity. The most integrated are scenario tasks. Six of these are available online (http://languagetesting.info/whatis/scenarios/list.php) as part of a package of multimedia resources to support apprenticeship-type learning programmes.

Scenarios were developed to simulate real-life assessment problems in which learners develop the craftsmanship of item and test development. In each case the stimulus is a set of multimedia materials that provide the setting for describing the need for a language test within a particular language use domain. Text and video are static, but news items about the use of tests within a domain, and links to relevant tests, are all updated on the fly every week. Learners take on the roles of professional test developers who have been hired to design an operational test for use in the domain. The outcome may be either a small number of tasks that would form part of a test, or a complete test.

The scenarios currently cover six areas: aviation, peacekeeping, call centres, medical communication, legal interpreting, and international student university admissions. I follow an eight-stage apprenticeship learning process, which I illustrate here with reference to the aviation language assessment scenario.

(a) Test Purpose and Issue Awareness: students first study the language standards of the International Civil Aviation Organization (ICAO), and consider the consequences of making decision errors within this context. Students show how the ICAO communication standards might guide test development. Key constructs relevant to aviation communication are identified and listed.
(b) Language Study: students study the restricted code of aviation language and then go to the website of the Aviation Safety Network to map communication failures against recorded accidents or near accidents. They are also presented with recordings of aviation communication where language contributed to accidents. In each case they are asked to identify the source and reason for miscommunication. Consideration is given to the use of non-standard language and its use when the restricted code is either violated or is not sufficient in unusual circumstances.
(c) A literature review to discover how others have approached this problem, as well as the social and ethical issues that come into play.
(d) Task design: students listen to and record live communications from an airport while tracking a plan on flightradar.com. The recording is used to develop a listening task for an aviation assessment battery. This may involve using the authentic recording,

or a scripted adaptation that the students write themselves and re-record for the test. Decisions at each stage of the audio design are recorded along with reasons for those decisions.

(e) Item types to assess identified constructs using the listening texts are designed in groups. Students record design decisions and reasons for the decisions taken. Item specifications are created.

(f) Qualitative Evaluation: students compare their tasks with those in existing tests used for the aviation industry. Where possible a prototype task is used with a language learner to practise protocol analysis leading to iterative task revision. Item specifications are evolved, each evolution is given a new number, and the rationale for each evolution is recorded.

(g) Quantitative Evaluation: It is normally extremely difficult for learners to prototype and pilot the new tasks in this domain, but in others (such as higher education) access to a small number of participants is often possible. Where it is not possible artificial datasets may be created for students to analyze using a range of tools such as the reliability or DIF tools in SPSS, or Excel spreadsheets for item analysis like those provided on the website (http://languagetesting.info/statistics/excel.html). Students are introduced to the software for quantitative analysis through practice under guidance in computer laboratories.

(h) Report Writing: writing a group or individual report on the process of developing, prototyping and piloting a sample task.

In integrated projects of this kind all elements of DISC are present. Control is exercised only through the provision of stimulus material from the website. Support is available from the expert tutor, but the role of the tutor is primarily to guide the students to make their own design choices as apprentice craftsmen, to justify those choices, and to evaluate them using the data that they collect. The involvement comes through the requirement to create and evaluate an artefact with a critical purpose in the real world that serves a social good; and to achieve this discussion and debate are essential. Just as important, the students learn that the process of test design and development is also a research process from beginning to end. This research is not possible without the integration of theory and practice, the careful weighing of social, ethical and societal demands, and weaving these into an effective and practical assessment tool.

7. Crafted Artefacts

The assessment artefact (step h) is not some isolated essay question, but the outcome of a process of design, building, and evaluation. The report produced in step h is therefore a radically different kind of output from those normally generated in postgraduate programmes, many of which have at best a tangential link to module content. Creativity in thinking about appropriate outputs for assessment is possible when the task characteristics and content are made explicit. In a recent student-driven activity, learners conducted research into the nature, design and intention of test accommodations for test takers with specific learning difficulties in their L1, when taking language tests in an L2. This is an important aspect of test design (sometimes called "universal design") that is frequently overlooked. The artefact produced at the end of the process was a podcast that problematized issues surrounding construct-irrelevant variance and made the research and potential solutions accessible to a wider audience. Following steps similar to those above, the final artefact integrated design issues with learning theory, accessibility and fairness, validity and reliability. The readers may judge the quality of this particular student artefact for themselves at http://languagetesting.info/features/test/accommodations.html, which is made available with the permission of the students. This is a prime example of genuine involvement in doing work for which they will take responsibility.

8. The Apprenticeship Classroom

The apprenticeship classroom for language teachers on an MA programme is primarily one of activity surrounding the content of a model of language assessment literacy delivered through tasks that follow the test design cycle. Some activities focus upon particular aspects of test design craft such as item writing and review, prototyping, or the analysis of pilot data. Other activities use the resources provided on the website languagetesting.info to support debate or group project work, such as the scenarios. Each task interweaves theoretical considerations, practical research, and test building. Students do this by creating small-scale tests and taking them through multiple steps in the design cycle, and then presenting the results to their peers in presentation sessions. Feedback is provided both by peers and the class tutor, which leads to iterative improvement. Assessment is closely linked to these activities. Writing up a scenario-based task design and analysis project as a research report is one option; but other options are possible, such as producing a broadcast quality podcast on some pressing assessment issue in the news, such as how language testing may be used to recruit health professionals to the health service while ensuring that patients are protected and can communicate with those who care for them in hospital. The apprenticeship classroom therefore looks and feels, as Peirce put it, like "doing" the work for real, experiencing the process, and taking responsibility for the outcome, even if it is in the service of completing an apprenticeship.

9. Conclusions

The LAL literature is replete with calls for the low literacy levels of stakeholder groups to be investigated and improved. What has not been addressed is how this is to be done. This chapter provides an initial articulation of one possible approach to LAL teaching and learning for language teachers on an MA programme, which is integrated with a theoretical model of LAL. It may go some way to encouraging teachers of language testing and assessment to consider how they use the resources available to them, including the website considered above.

It is suggested that the apprenticeship model is particularly appropriate for language teachers and others whose careers will require them to design, develop and administer language tests. The quotation from Peirce above makes clear that the pedagogic model is underpinned by a pragmatic philosophy that insists upon the interconnection between knowledge, learning, and experience. The "doing" as an apprentice is also an inherently social activity that requires collaboration with others. In education generally Dewey "…urged workers to assess the quality of their work in terms of shared experience, collective trial and error" (Sennett 2009, 288), which is fundamental to an apprenticeship model of LAL pedagogy. The apprenticeship model also resonates with the view that LAL activities are useful for the professional development of practising teachers, many of whom may also be undertaking MA programmes (Koh 2011; Fulcher 2017). It has long been recognized that engaging in research and collaborative design enhances professional practice: the link between doing and growing as a practitioner is an essential feature of the teacher research movement (Hopkins 2008, 38). Through apprenticeship-like tasks teachers acquire theory and knowledge as well as skills, becoming autonomous professionals with the capacity to make informed decisions (Goodenough 2010).

I make no claim that other pedagogic models may prove more useful, particularly for other stakeholder groups. However, I do contend that this is an explicit, deliberate pedagogy, as Brown (2011, 145) has called it, which to date has not been articulated in the literature. That LAL research has reached the stage of operationalization in pedagogies based on sound theory and research is also a sign of maturity within the discipline. Only in established self-confident branches of learning do current practitioners consider how best to serve the needs of the next generation. Interest in the theory and pedagogy of LAL is indicative of a profoundly optimistic field of endeavour.

The primary purpose of addressing the question "How do we do it?" is to achieve the assessment literate society that Stiggins envisioned. In turn, this requires what Fred Davidson (personal communication) calls "testing citizenship". This term echoes Wenger's notion of social participation, with specific reference to "the ethics of how we invest our identities as we travel through the landscape [of practice]" (Wenger, 92), bringing a relevant understanding of assessment to the decisions that we are required to make. This chapter makes one small contribution towards realizing such a goal through LAL practice.

6. References

Berry, Vivien, and Susan Sheehan. 2017. "Exploring teachers' Language Assessment Literacy: A Social Constructivist Approach to Understanding Effective Practices." Paper presented at the 6th ALTE Conference, Bologna. Online: http://eprints.hud.ac.uk/31898/.

Black, Paul, and Dylan Wiliam. 2009. "Developing the Theory of Formative Assessment." *Educational Assessment, Evaluation and Accountability* 1(1): 5-31.

Bracey, Gerald. 2000. *Thinking about Tests and Testing: A Short Primer in "Assessment Literacy"*. Washington DC: American Youth Policy Forum. ERIC Document No. ED 445 096.

Brookhart, Susan. "Educational Assessment Knowledge and Skills for Teachers." *Educational Measurement: Issues and Practice* 30(1): 3-12.

Brown, A. 2011. Review of Language Testing and Assessment: An Advanced Resource Book. *Language Testing* 28(1): 145-148.

Candlin, Christopher. 1987. "Towards Task-based Language Learning." In *Language Learning Tasks,* edited by Christopher Candlin, and D. F. Murphy, 5-22. Lancaster Practical Papers in English Language Education Vol. 7. London: Prentice Hall.

Conole, Grainne, Martin Dyke, Martin Oliver, and Jane Seale. 2004. "Mapping Pedagogy and Tools for Effective Learning Design." *Computers & Education* 43(1-2): 17-33.

Coomey, Marion, and John Stephenson. 2001. "Online Learning: it is all about Dialogue, Involvement, Support and Control–According to the Research." *Teaching and Learning Online: Pedagogies for New Technologies*, 37-52.

Crusan, Deborah, Lia Plakans, and Atta Gebril. 2016. "Writing Assessment Literacy: Surveying Second Language Teachers' Knowledge, Beliefs, and Practices." *Assessing Writing* 28: 43-56.

Davies, A. 2008. "Textbook Trends in Teaching Language Testing." *Language Testing* 25(3): 327-348.

DeLuca, Christopher, and Don A. Klinger. 2010. "Assessment Literacy Development: Identifying Gaps in Teacher Candidates' Learning." *Assessment in Education: Principles, Policy & Practice* 17(4): 419-438.

Dewey, John. 1923. *Democracy and Education: An Introduction to the Philosophy of Education*. Macmillan. Project Guttenberg Edition (http://www.gutenberg.org/files/852/852-h/852-h.htm).

Deygers, Bart, and Margaret E. Malone. 2019. "Language Assessment Literacy in University Admission Policies, or the Dialogue that isn't." *Language Testing*, https://doi.org/10.1177/0265532219826390.

Fulcher, Glenn, and Fred Davidson. 2007. *Language Testing and Assessment: An Advanced Resource Book*. Routledge.

Fulcher, Glenn. 2012. "Assessment Literacy for the Language Classroom." *Language Assessment Quarterly* 9(2): 113-132.

—. 2013. *Practical Language Testing*. Routledge.

—. 2015. *Re-examining Language Testing: A Philosophical and Social Inquiry*. Routledge.

—. 2017. "Cultivating Language Assessment Literacy as Collaborative CPD." Paper presented at the BALEAP Conference, Bristol.

Goodnough, Karen. 2010. "The Role of Action Research in Transforming Teacher Identity: Modes of Belonging and Ecological Perspectives." *Educational Action Research* 18(2): 167-182.

Harding, Luke, and B. Benjamin Kremmel. 2016. "Teacher Assessment Literacy and Professional Development." In *Handbook of Second Language Assessment*, edited by Dina Tsagari, and Jayanti Banerjee, 413-427. Berlin: De Gruyter.

Hasselgreen, Angela, Cecile Carlsen, and Hildegunn Helness. 2004. "European Survey of Language Testing and Assessment Needs. Part One: General Findings." *Online: www. ealta. eu. org/resources [23.07. 2011]*.

Hill, Kathryn. 2017. "Understanding Classroom-based Assessment Practices: A Precondition for Teacher Assessment Literacy." *Papers in Language Testing and Assessment* 6(1): 1-17.

Hopkins, David, Bruce Joyce, and Emily Calhoun. 2008. *A Teacher's Guide to Classroom Research*. Open University Press.

Huai, Nan, Jeffery P. Braden, Jennifer L. White, and Stephen N. Elliott. 2006. "Effect of an Internet-based Professional Development Program on Teachers' Assessment Literacy for All Students." *Teacher Education and Special Education* 29(4): 244-260.

Huhta, Ari, Tuija Hirvalä, and Jayanti Banerjee. 2005. "European Survey of Language Testing and Assessment Needs. Part 2: Regional Findings." online: http://users.jyu.fi/~huhta/ENLTA2/First_page.html.

Koh, Kim H. 2011. "Improving Teachers' Assessment Literacy through Professional Development." *Teaching Education* 22(3): 255-276.

Lam, Ricky. 2015. "Language Assessment Training in Hong Kong: Implications for Language Assessment Literacy." *Language Testing* 32(2): 169-197.

Lester, Richard K., and Michael J. Piore. 2009. *Innovation—The Missing Dimension*. Harvard University Press.

Manning, Anthony. 2016. *Assessing EAP: Theory and Practice in Assessment Literacy*. Garnet.

Ofra Inbar-Laurie. 2016. "Language Assessment Literacy." In *Language Testing and Assessment*. Encyclopedia of Language and Education, edited by Elana Shohamy. New York: Springer.

Peirce, Charles Sanders. 1878/1958. "Research and Teaching in Physics." In *Values in a University of Chance. Selected Writings of Charles S. Peirce*, edited by Philip Wiener, 325-330. New York: Doubleday and Co.

Pica, Teresa, Ruth Kanagy, and Joseph Faldun. 1993. "Choosing and Using Communication Tasks for Second Language Instruction." In *Tasks and Language Learning: Integrating Theory and Practice,* edited by Graham Crookes, and Susan Gass, 9-34. London: Multilingual Matters.

—. 1993. "Choosing and Using Communication Tasks for Second Language Instruction." *Multilingual Matters*, 9-34.

Pill, John, and Luke Harding. 2013. "Defining the Language Assessment Literacy Gap: Evidence from a Parliamentary Inquiry." *Language Testing* 30(3): 381-402.

Popham, James. 2009. "Assessment Literacy for Teachers: Faddish or Fundamental?" *Theory into Practice* 48(1): 4-11.

—. 2011. "Assessment Literacy Overlooked: A Teacher Educator's Confession." *The Teacher Educator* 46(4): 265-273.

Read, John. 2011. “Review of Practical Language Testing.” *Language Testing* 28(2): 302-304.

Sennett, Richard. 2009. *The craftsman*. Penguin.

Shaffer, David Williamson. 1993. “Assessment Literacy for Teachers.” *Theory into Practice* 32(1): 118-126.

Sheehan, Susan, and Sonia Munro. 2017. "Assessment: Attitudes, Practices and Needs." British Council ELT Research Papers 17.08.

Smith, Calvin Douglas, Kate Worsfold, Lynda Davies, Ron Fisher, and Ruth McPhail. 2013. "Assessment Literacy and Student Learning: The Case for Explicitly Developing Students ‘Assessment Literacy’." *Assessment & Evaluation in Higher Education* 38(1): 44-60.

Stiggins, Richard J. 1995. "Assessment Literacy for the 21st Century." *Phi Delta Kappan* 77(3): 238.

Taylor, Lynda. 2013. "Communicating the Theory, Practice and Principles of Language Testing to Test Stakeholders: Some Reflections." *Language Testing* 30(3): 403-412.

Tsagari, Dina, and Karin Vogt. 2017. "Assessment Literacy of Foreign Language Teachers around Europe: Research, Challenges and Future Prospects." *Papers in Language Testing and Assessment* 6(1): 41-63.

Vogt, Karin, and Dina Tsagari. 2014. "Assessment Literacy of Foreign Language Teachers: Findings of a European Study." *Language Assessment Quarterly* 11(4): 374-402.

Wenger, Etienne. 2010. "Communities of Practice and Social Learning Systems: The Career of a Concept." In *Social Learning Systems and Communities of Practice*, 179-198. Springer, London.

Xu, Yueting, and Gavin T. L. Brown. 2016. "Teacher Assessment Literacy in Practice: A Reconceptualization." *Teaching and Teacher Education* 58: 149-162.

—. 2017. "University English Teacher Assessment Literacy: A Survey-test Report from China." *Papers in Language Testing and Assessment* 6(1): 133-158.

CHAPTER TWO

FINDING OUT WHAT LEARNERS KNOW–AND…? REFLECTIONS ON TEACHERS' LANGUAGE ASSESSMENT LITERACY

GUDRUN ERICKSON

The current text focuses on assessment as an obvious aspect of [language] education, inseparable from learning and teaching. Together, the three concepts form a pedagogical trinity, fundamental to language teachers' professional competence. Different aspects of language assessment are concretized and exemplified, thereby offering an outline of what can be referred to as Language Assessment Literacy (LAL). First, this means reflection on the definition of assessment, how it relates to teaching, feedback and testing, and how it should be approached in communication with students and other stakeholders. Secondly, it involves aspects of purpose(s), i.e. the reason(s) why assessment is taking place, the answer to which will obviously affect the methods chosen. Context also plays a role in this. Furthermore, the "knowledge object" of the assessment, the construct, has to be defined, as will the procedures and tasks developed and used, including ways to interpret, present and discuss results. Also, the question of agency needs attention, that is who does what, not least in what ways students are–or are not–actively involved in assessment processes. Finally, questions of use and consequences, the way teachers act on observations made and the results obtained are crucial, for students as well as for teachers, schools and language education at large. The above aspects are also briefly discussed in relation to the fundamental concepts of validity and reliability–what they are and how they may be used in teachers' work. Finally, some references are made to international sources and resources for the enhancement of language teachers' assessment literacy.

1. Introduction

Deciding to become a teacher may have a number of reasons, for example a wish to work closely with people, an ambition to assist these people in their learning and growing, love for a certain subject or group of subjects, and/or the prospect of working within a field offering a potential for flexibility, creativity and personal development. Whatever the reason(s) may be, however, the decision always means embarking on a long journey together with a large number of companions, whose learning and development you will not only follow for a considerable amount of time, but also strive to support and affect in as many positive ways as possible. Variation in these groups will be obvious at group as well as individual levels and related to widely different reasons. To handle this variety, and to build on it for the benefit of all students, is one of the main and most positive challenges of the profession.

Another general characteristic of the teaching profession is the need to handle various aspects of learning, teaching and assessment, which requires specific competences–theoretical as well as practical–within the different areas as well as regarding the combination of the three.

As indicated in the title, this text will focus on assessment, in particular the competences that teachers of languages need to handle this broad concept. First of all, this is based on the premise that there is a distinct relationship between learning, teaching and assessment, made explicit, for example in the Common European Framework of Reference for Languages: Learning, Teaching and Assessment (Council of Europe 2001). In this, the connection between learning and teaching is seldom disputed in any basic sense, although there is no automatic or linear relationship between the two. Furthermore, and not focused upon quite as often, is the fact that there are clear similarities between teaching and assessment, for example regarding relevance, breadth, clarity, sensitivity, fairness, and respect. However, and not to be ignored, there are also obvious differences, which will be touched upon later. Finally, it needs to be remembered, that the most obvious relationship between learning, teaching and assessment is that teaching and assessment have the same ultimate goal, namely to promote learning in all sorts of ways and to use methods that make students develop in as many positive ways as possible, both regarding competence and confidence. Due to all this, and risking stating the obvious, assessment is an inevitable aspect of teachers' professional competence: when you are genuinely interested in students learning something that you yourself find absolutely essential, it would be phenomenally strange not to be interested–and skilled–in finding out if they have learnt, what and how well they have learnt, and what they themselves consider important in relation to their own continued work and development.

Given the aim of the chapter, namely to outline and reflect on the concept of language assessment literacy (Fulcher 2012; Inbar-Lourie 2008; Stiggins 1995; Tsagari and Vogt 2017), a number of basic issues will be touched upon–issues that can be used for and by teachers, both for pre- and in-service educational purposes, and in planning different assessment activities. How is language competence defined, and what does the CEFR expression "action oriented approach" imply, in general and for assessment? What are the different purposes of assessment? How can we construct tasks and items that offer as many students as at all possible the chance to show what they *can* do with their language, not making them feel that the opposite is in constant focus? When, and how often, is assessment taking place and what do we actually mean by the concept as such? In what way is our use and definition of the word perceived by students and others? Who is taking part in the different phases of assessment activities, and last but definitely not least: How are results used and what effects are there for students as well as for teachers – and further, at a structural level? In the last part of the chapter some concrete examples will be given of sources and resources regarding Language Assessment Literacy.

2. Perceptions and functions

It is sometimes claimed, that since almost everybody has been to school, everybody feels entitled to have definite opinions about school. This often concerns issues related to assessment and grading, areas where strong opinions are often expressed, probably based on personal experiences. These memories may be positive or negative but quite often they focus on perceived injustice and anxiety in a way that often makes assessment appear as something negative, something that has to do with power and lack of power, and that much too often seems to be intended to make people show what they *cannot do* instead of the opposite, what they *can do* (Erickson and Sebestyén 2018). However, a certain change in this has been noticeable during the last few decades, when definitions and concrete descriptions of assessment having a positive pedagogical potential have emerged, based on theory as well as documented experiences. Concepts like "formative assessment" and "assessment for learning" are often used, and assessment is more and more often regarded and treated as a natural component of education at large (Gardner 2012).

The pedagogical function of assessment–assessment *for* learning–seems obvious and is defined in the following way by the Assessment

Reform Group (2002)–a group of researchers in varying disciplines related to education, active in the UK from the late 1980s to 2010:

> *Assessment for Learning is the process of seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning, where they need to go and how best to get there.*

This definition is fundamental to the use of assessment in education: to identify individual knowledge profiles with their different strengths and areas of necessary development, to clarify goals, and to discuss the best way(s) to achieve these goals. Furthermore, it is worth emphasizing that this is equally important to students and to teachers.

It is, however, essential to remember that assessment also has a central role in the duty of schools to ensure quality and contribute to equity, hence, to work with assessment *of* learning. There is by no means a contradiction between this and the pedagogical function; rather, the two should be seen as complementing each other and creating a basis for planning the way forward.

There are a number of underlying, generic principles regarding the quality of assessment, regardless of the aim and function. These principles can be phrased in different ways but are always, one way or the other, related to clarity, validity, reliability and respect. Students should never have to guess what the goals of their learning are or how these goals are going to be assessed. Further, assessments need to be relevant in relation to what they are intended to capture, and it should be possible to trust the results as far as possible as indicators of students' knowledge and competence. And finally, maybe the most fundamental principle of them all: assessments of different kinds should be ethically developed and implemented, meaning that students are shown the respect they indeed deserve in their learning and development. (The above-mentioned concepts and principles will be further discussed later in the text.)

3. Fundamental issues to be addressed

Language assessment literacy requires a clear *DEFINITION* of what assessment is, not least regarding its role in education at different levels. As has been pointed out already, this means handling aspects of learning, teaching and assessment and the different relationships between the three, but also being aware of the dual role of assessment in enhancing learning as well as fairness and equity. Furthermore, there are certain basic questions to be addressed, preferably together with colleagues and students, to define and optimize processes and instruments developed to capture both stages of development and levels of competence in students' language proficiency (Takala, Erickson, Gustafsson and Figueras 2016).

The first question concerns *PURPOSE* and has to be dealt with early in the process. This is because the reasons to assess influence both the type of tasks or procedures developed and the way in which they are implemented. Furthermore, *CONTEXT* may be seen as an aspect of this: there is a considerable difference, for example, between international surveys, national tests related to grading, teachers' continuous feedback and assessment in classrooms, and students' own evaluations of what has been learnt (Takala et al. 2016). How to handle this variety, albeit based on the same fundamental principles of quality, is something that benefits greatly from a stable conceptual basis as well as from communication and collaboration. Also, within the framework of classroom assessment, different purposes indeed have an impact, for example regarding issues of focusing, format and analyses. In this, formative and summative aspects are included, as well as the potential to concretize what language learning and knowing are all about.

Clarifying, for example, a functional way of defining language competence is related to the necessary interpretations and implementations that have to be made of the *CONSTRUCT,* that is what students are expected to learn and teachers to teach and assess. This is specified in curricula of different kinds, be they local, national or international. Current documents of this type are most often characterized by a communicative and functional definition of language that can be traced back to Dell Hymes' (1972) groundbreaking work and clarifications and elaborations made by Canale and Swain (1980) and others. In this, communicative competence is regarded as an overall competence comprising a number of sub-competences, in the CEFR defined as linguistic, sociolinguistic and pragmatic competences. The core aspect of this is to use language in communication with others, as far as possible about authentic/"real" content. The interpretation of this is sometimes hard, regarding for example means and goals–what is needed to make language work in active use. There are several ways to approach and explain this, not least to students. One way is to make use of the common metaphor about building a house, where in this case building blocks, mortar and surface materials are self-evident, but the house itself with its quality, functionality and comfort is the obvious goal. In other words: you cannot learn and further develop your language without words and phrases, grammar and syntax, spelling, pronunciation and prosody, but it is the functioning language at large that is in focus, that you both can and want to use for communication in various situations, for numerous purposes and with different people. The building blocks are of course indispensable, as is what is used to make them stay together and be sustainable, but they are not the goals per se of your building venture (Erickson 2016). From an assessment perspective, this obviously means that you need to be careful not to let the means take over and dominate over the goals, that is to allow elements like vocabulary, grammar and spelling be the main target instead of language use, being the goal of language learning.

The question about *METHODS* is multi-facetted and encompasses a number of different aspects of assessment: the development of tasks, including the issue of modality and format, principles for handling administration and rating, decisions about analyses and standards for different qualitative levels and, importantly, the development of ways to present results that can achieve what Sadler (1989) refers to as feeding back and feeding forward. This requires continuous analyses of purposes and constructs, that is what is intended to be assessed and why, of results–for all students and subgroups of students–and of users' different opinions and reactions. All of this is related to the basic principles mentioned earlier and implies that the ways in which students are expected to respond and act should be varied. This is in line with what experience and research show, namely that people learn in different ways, both regarding methods and time. Consequently, it seems highly reasonable that they should also be given the opportunity to show their knowing in different ways. However, what is most adequate varies to some extent with purpose and context: some formats, for example multiple choice items, are most fit for tests where solid pretesting procedures are used to optimize quality, whereas more dynamic types of tasks may be easier to administer in small-scale contexts. What can also be mentioned is that a variation of formats and types of tasks seems to be appreciated by pupils, shown for example in a large, international study of students and teachers' views on language testing and assessment (Erickson and Gustafsson 2005).

Another aspect of the methodological issue is how to achieve as much stability and reliability in rating as possible, thereby enhancing quality. One aspect of this is the question of analytic and/or holistic approaches, that is whether global ratings or the rating of different aspects separately, with or without final aggregation, is preferred. Research does not give a definite answer to the question of whether one or the other method is superior in quality; it is rather highly dependent on the purpose of the intended assessment. This is also related to the use of what is sometimes referred to as rubrics, sometimes as matrices, whose purpose is to clarify the content presented in the curriculum by breaking it down into smaller units that can be handled separately. In this case, as in many others, needs and decisions must be analyzed in relation to purpose and intended effects, in which intended clarity has to be considered in relation to the risk of fragmentation. In line with this, Stobart (2012) warns about making the objectives *"so detailed that they encourage impoverished learning"* (p. 238), and Torrance (2007) talks about the risks of creating situations in which *"assessment procedures and practices come completely to dominate the learning experience, and 'criteria compliance' comes to replace 'learning'"* (p. 282). Hence, the good ambition to clarify what is to be learnt–and assessed–must not lead to the narrowing of the curriculum, or phrased differently, a pedagogical practice where what is easily measurable becomes the most important (Erickson and Åberg-Bengtsson 2012). Following the principles of constructive alignment (Biggs 1996), what is assessed must clearly be connected to the content standards, or learning objectives, expressed in the curriculum–that is, to the description given of the construct. Moreover, it is worth considering the implications of the fact that assessment is not additive in the sense that the sum of the different parts is always identical to the whole; there are very often additional and interactive factors that cannot be detected in easy-to-read and tick-off descriptions. It is in capturing these linking features in language use, to be able to combine the specific with the general, that teachers' overall analyses and professional judgement play a central role in assessment and thereby constitute an essential aspect of LAL.

The question about *TIMING* encompasses when and how often assessments are made and also how much time is spent on activities related to assessment. In this, it is important to be clear about purposes, not least to clarify to students how you distinguish between regular communication about ongoing learning activities, and assessment that is documented for summative purposes. This is indeed a demanding balancing act: on the one hand, you do not want to base your assessment decisions only on a few occasions that will then, most probably, be perceived as stressful by students; on the other hand, the opposite also has a number of risks. To tell your students, with the best of intentions, that they do not have to worry about tests, because you assess continuously, may not be the most calming or adequate message for everybody (Erickson 2016). What the teacher wants to say is most probably that students can demonstrate their competences in a variety of situations, not only in formal tests, but some students are likely to hear something else as well, namely that signs of *not* having understood or *not* being able to do some things will be documented as well. Messages of this type are ambiguous, in the sense that the more anxious students–if the worst comes to the worst–may feel that they are under constant scrutiny and should therefore be quiet, not revealing their possible misunderstanding or lack of understanding. It is of course extra serious in the case of languages, if teachers' words and expressions seem to signal that the self-evident and necessary search for understanding – sometimes short, sometimes long – will be assessed as such, when it is in fact an absolutely natural part of learning that needs to be given space and time. When considering this, it is interesting to note that the importance of creating reasonable clarity regarding the definition and timing of assessment is commented on in the CEFR (2001), where it is stated that continuous assessment *"can, if taken to an extreme, turn life into one long never-ending test for the learner and a bureaucratic nightmare for the teacher"* (p. 185).

Aspects of *AGENCY* are essential in all discussions about assessment and encompass issues concerned with the subject-object relationship, power, responsibilities, collaboration, and respect. First of all, it needs to be emphasized, that if you as a teacher feel that you are the only one responsible for the assessments needed and required, the burden will be very heavy and the outcome sensitive to several factors related to quality in a broad sense. Therefore, increasing the number and types of participants is highly recommendable. This means, first of all, to let students into the process by using sensibly both self- and peer assessment models to gain perspectives and angles you traditionally as a single teacher do not have access to. True, students do not have a formal role in assessment decisions and they obviously lack the kind of competence that teachers possess; however, they have insights and experiences that are quite unique and that help teachers make more adequate choices when planning both instruction and assessment. Furthermore, collaboration with colleagues concerning planning as well as use and analyses of assessment activities and results has the potential of providing invaluable support. In addition, opinions "from the outside", be they in the form of observations and discussions or results from assessment activities, add a widening of the field of view that is highly beneficial for assessment quality. Undoubtedly, more eyes looking and ears listening, and more voices being heard are of great value to learning and teaching as well as to assessment. This relates to the basic principles mentioned before, namely clarity, validity, reliability and respect, and increases the possibilities of making adequate observations and using them well.

Regarding the agency of the students–their activities and collaboration in assessment–much inspiration and many examples can be found in the CEFR as well as in the European Language Portfolio with its many national versions, which both in a systematic way focus on students' language awareness and active participation in their own learning–something that can be applied in classroom contexts as well as in more large-scale contexts (Little 2009; Little and Erickson 2015). Finally, in relation to the issue of agency, what should be self-evident all the same needs to be pointed out, especially since words and traditional expressions sometimes express something different: assessment is about establishing, as well as possible, levels of competence in relation to a defined construct, not about evaluating human beings. Consequently, it is crucial to be careful about choosing and using the right object when talking or writing about what is done. Bachman (1990) expresses this in the following way: *"Whatever attributes or abilities we measure, it is important to understand that it is these attributes or abilities and not the persons themselves that we are measuring"* (p. 20).

Finally, and indeed at the heart of the matter of the quality of assessment, issues focusing on the *USE AND CONSEQUENCES* of the results obtained have to be taken into account in all stages of the assessment process. This also relates to the old English proverb stating that "the proof of the pudding is in the eating", meaning that nothing can really be considered successful, if it has not been implemented in an adequate way, found useful for its purpose, and generated beneficial effects for students' learning and teachers' teaching. Furthermore, a structural aspect can be added to the levels where use and consequences are in focus. Shohamy (2001) emphasizes that assessment is by no means an isolated phenomenon but clearly integrated in social, educational and political contexts. She also applies an ethical perspective and warns about the direct or indirect execution of power through testing and assessment.

All the issues mentioned and commented on above, not least the final one focusing on use and consequences, are related to the quality of assessment and thereby to the fundamental concepts that will be focused upon in the following section of the text.

4. Fundamental concepts

Two of the most fundamental concepts related to assessment are *validity* and *reliability,* which have already been mentioned in connection with the generic principles introducing the section on issues to be addressed. Expressed in a basic way, validity can be said to deal with what is *right* regarding content, use and consequences, and reliability with *"righteousness",* or fairness in a broad perspective. Very much of what has been discussed in connection with the issues in the previous section is embedded in these two concepts.

Validity

For assessment to fulfil its function, it needs to be valid, or useful and relevant, for its purpose. This means that it should capture, as far as possible, what has been defined as the construct in a specific learning situation and do it in the best way possible from a measurement as well as a pedagogical point of view. Traditionally, validity has been defined as just this, that you should measure what you intend to measure, nothing else, and do it as reliably as possible. This ambition certainly remains but it has been considerably expanded into focusing on the use of results–the inferences and decisions made and the actions taken on the basis of what has been seen (Messick 1989; Moss, Girard and Haniford 2006). Although somewhat controversial, consequences have also been added to the definition of the concept, defined by Messick (1989, 13) in the following way: *"Broadly speaking, then, validity is an inductive summary of both the existing evidence for and the potential consequences of score interpretation and use"*. "Consequences" are here used in a broad sense, comprising individual as well as pedagogical and structural aspects, including the impact that assessment and tests may have both on students' views on their own competences and teachers' professional practices, the latter referred to as *washback effects* (Green 2013; Messick 1996). The expansion of the concept of validity thus means that both those who develop instruments and procedures and those who implement and use them have a responsibility for what happens, or may happen, based on the results. This requires continuous quality work–often referred to as validation (Kane 2006).

A number of threats to validity can be identified, often summarized in two categories, namely *construct under-representation* and *construct-irrelevant variance* (Messick 1989). Under-representation means that too little of the construct is included in assessments, where instead what is perhaps easier to capture and measure, is given much attention. An example of this may be taken from the European Survey on Language Competences (European Commission 2012), in which oral language proficiency was not included in spite of being a distinct and important part of the construct, based on the CEFR (the reason, according to the European Commission 2012: 4, being that "Speaking was considered logistically difficult"). The other category, *construct-irrelevant variance,* occurs when something other than the intended construct is in focus, for example when assessment requires writing to a large extent, also in cases where writing is not in focus (for example when assessing receptive competence). The examples given here are fairly obvious, but of course there are also cases that are more difficult to detect. However, it has been shown that even fairly young students intuitively describe and define both under-representation and irrelevant variance, when asked to characterize what they feel may be negative about assessment. Two examples, taken from data collected by Erickson and Gustafsson for their 2005 study are the following:

A bad test/assessment is the one which is only about grammar, because if a person knows grammar well, it doesn't mean he/she can speak the language as well and communication is the most important thing in language study…

… a bad test is when we just have a little time to do it on. Because then you have to stress thru the test and maybe you cannot show how much you can…

Reliability

A prerequisite for the claim that a certain assessment is valid is that the tasks and procedures used to generate results are reliable, that is, they can be trusted in a broad sense. Reliability can be defined and estimated in different ways, but basically it means that the results obtained shall reflect "true" competence as precisely as at all possible (AERA, AEA and NCME 2014). This means, for example, that different errors of measurement have to be identified and eliminated or minimized, errors of both a random and systematic kind. Further, the amount of data used to draw conclusions and make decisions needs to be large enough. At a concrete level, it is crucial that students actually understand what they are supposed to do and that the scoring guides for teachers are clear and extensive enough to support consistent ratings. The latter is especially important in open-ended tasks, where students use their language actively in oral and written production and interaction.

The concepts of validity and reliability are closely related, and as already mentioned, validity requires reasonable reliability. However, the reverse does not apply: an assessment can demonstrate high reliability without being valid for its purpose. It is also worth mentioning that reliability is sometimes regarded as an aspect of validity, for example referred to as scoring validity by Weir (2005). Validity is then seen as the overarching concept, in some ways synonymous to quality.

Everyday validity and reliability

The concepts of validity and reliability are sometimes perceived as fairly technical and theoretical, best suited for academic analyses of different kinds but hardly relevant in concrete pedagogical contexts. Albeit partly understandable, given the way the concepts are often talked about–or not talked about–this attitude needs to be challenged, since the concepts are indeed highly useful both for pre- and in-service training purposes, especially regarding the planning and analysis of assessment. Due to this, it seems highly recommendable that teachers, on their own and together, ask continuous questions of the following kinds:

- Do my/our assessments reflect the view of language and language use expressed in the curricula?
- What is the balance between assessments focusing on speaking, writing and "doing"?
- Is the construct covered in a reasonable way, or are some aspects given more or less attention? If so, why?
- How varied are the tasks I/we use regarding content and format?
- To what extent and in what ways are students active in assessment procedures?
- Do I/we have sufficient evidence for making decisions about students' levels of competence?
- To what extent do I/we discuss assessment issues with students and colleagues, and how is this organized?
- How much do I know about my level of agreement with colleagues (and myself) regarding ratings?

These questions and others clearly reflect fundamental principles and concepts related to language assessment, but they are at the same time highly concrete and useful in successive, collegial work focusing on developing and optimizing assessment practices.

5. Examples of sources and resources regarding Language Assessment Literacy

The vast majority of all assessment takes place in classrooms. Hence, teachers form a highly important group of actors and stakeholders and also, together with their students, a category that may be seen as the real protagonists in testing and assessment. Furthermore, and in light of what has been said earlier about validity, teachers are also a crucial category regarding what is referred to as consequential validity (Messick 1989; 1996), being deeply involved in the effects of different types of assessments both on students' learning and teachers' teaching. To handle this in the best possible ways, teachers need to acquire and further develop their Language Assessment Literacy, that is multifaceted knowledge in and about assessment as such, but also as a phenomenon embedded in an educational context, thus requiring a distinct breadth and depth of knowing, including policy issues and framework factors of different kinds.

Awareness of and interest in LAL have grown distinctly during the last few decades, partly due to the increasing focus on formative aspects of assessment, assessment for learning, learning-oriented assessment, dynamic assessment, etc. This is obviously positive, but it also requires plenty of knowledge and experience to be able to distinguish between what is more or less useful for different purposes and contexts. Hence, LAL is more important than ever.

It would hardly be possible, and probably not very useful, to try to refer to a large number of sources and resources regarding Language Assessment Literacy. There is an increasing body of research with accompanying academic articles as well as materials connected to specific tests–all of this with varying accessibility, in a concrete as well as a metaphoric sense. What will be provided in the following is therefore a brief mention of four different, international sources that are easily available for teachers and others interested in concrete materials and what could be called "applied reasoning" with the aim of enhancing LAL. The four sources are The Council of Europe, The European Centre for Modern Languages (ECML), The European Association for Language Testing and Assessment (EALTA), and the Teachers' Assessment Literacy Enhancement project (TALE), co-funded by the European Commission.

THE COUNCIL OF EUROPE (CoE) website includes a large number of materials related to language policy and language education, very useful for a number of purposes, one of which is the enhancement of Language Assessment Literacy. In this, information and commentaries on the CEFR and its Companion Volume (Council of Europe 2018) are obviously a valuable resource, which is certainly also true for The European Language Portfolio (ELP) and its accompanying materials (https://www.coe.int/en/web/portfolio). The ELP is the property of the learner and was created to support "the development of learner autonomy, plurilingualism and intercultural awareness and competence" and also "to allow users to record their language learning achievements and their experience of learning and using languages". It has three components: a language passport, a language biography and a dossier for use by students to record and reflect on their development and experiences. It has been translated and adapted to different educational sectors in many countries (see the CoE website). One of the core aspects of the ELP is to enhance students' awareness of language learning and the ability to evaluate their own levels of competence. This has an obvious relationship to the principles of transparency and validity as well as to the question of agency mentioned previously and it corresponds well with the ethical dimension of consequential validity. The ELP thereby offers additional ways for language learning, teaching and assessment, and thus also contributes to teachers' Language Assessment Literacy.

Work on language policy is only one part–albeit essential–of the activities of the Council of Europe, but there is also "A Centre to Promote Quality Language Education in Europe", namely THE EUROPEAN CENTRE FOR MODERN LANGUAGES (ECML) in Graz, financed by a number of member states of the CoE. The ECML has a close relationship with, and a complementary function to, the CoE. The Strategic Objectives of the Centre are to help its member states implement effective language teaching policies by:

- focusing on the practice of the learning and teaching of languages,
- promoting dialogue and exchange among those active in the field,
- training multipliers, and
- supporting programme-related networks and research projects.

During its more than 20 years, the ECML has conducted a large number of projects in active collaboration with its member states. Information about these projects, some of them focusing on assessment, can be found on the website of the Centre (http://www.ecml.at), where different types of materials are also available.

There are a number of international organizations focusing on language testing and assessment. One in particular, THE EUROPEAN ASSOCIATION FOR LANGUAGE TESTING AND ASSESSMENT (EALTA), founded in 2004, has a broad approach to testing and assessment and also an explicit ambition to include a broad group of members interested in language assessment. In the preparatory work preceding EALTA, a number of studies distinctly related to the issue of Language Assessment Literacy were undertaken, two of them focusing on Language Testing and Assessment Needs (Hasselgreen, Carlsen and Helness 2004; Huhta, Hirvelä and Banerjee 2005) and one on European Students' and Teachers' Views on Language Testing and Assessment (Erickson and Gustafsson 2005). These reports are available on the association website (http://www.ealta.eu.org/resources.htm). In another section of this website, EALTA's Guidelines for Good Practice in Language Testing and Assessment can be downloaded (http://www.ealta.eu.org/guidelines.htm). They are available in 35 language versions and include a general introduction to the topic, also stating the inclusive mission statement of the association, and after that four separate sections, three approaching language testing and assessment in teacher education, classrooms, and national or institutional testing units or centres, and a final section focusing on linkage to the Common European Framework.

Finally, in this section on sources and resources regarding LAL, the TALE project (2015-2018) is well worth mentioning (http://www.taleproject.eu/). TALE is–or rather was–a project co-funded by the European Commission within the Erasmus+ Programme, with a project team comprising members from Cyprus (the coordinating university), Germany, Greece, Hungary, Norway and the UK. In the project, an online, self-study training course was developed, intended for use by pre- and in-service teachers of English and other foreign languages, as well as by teacher educators to raise awareness and levels of LAL. The online course is accompanied by a handbook, which includes eight chapters plus a glossary, all of which covers a number of theoretical and practical issues. The eight chapters, including short videos, focus on the following:

- The ABCs of Assessment
- Assessing Reading Skills
- Assessing Writing Skills
- Assessing Listening Skills
- Assessing Speaking Skills
- Providing Feedback
- Alternatives in Assessment
- Test Impact.

The course is based on distance learning methodology and participants are expected to adapt the materials to their own contexts. Reflection plays an important role. All in all, the materials offer excellent opportunities to learn and develop further, individually and in interaction with others, within the field of language testing and assessment.

6. Concluding remarks

Assessment is an aspect of education that often evokes strong feelings and triggers opinions of varying kinds. This may seem strange, since finding out if, what and how, and also how well students have learnt can easily be regarded as a highly natural and recommendable activity in education to help students as well as teachers to plan ahead and optimize learning and teaching. However, the activity also has the opposite potential, namely, to highlight most of all what has not been learnt and what may be perceived as overwhelmingly difficult, thereby risking making students feel small instead of big, to shrink rather than grow. Learning to handle assessment in the best possible ways, with flexibility and creativity, sensitivity and respect is therefore a crucial aspect of teachers' professional competence, requiring ample knowledge about theories within the field as well as an acquaintance with good practices and positive examples. Assessment Literacy may be seen as a bridge between these two sources of competence, both indispensable and depending on each other.

7. References

AERA, AEA, and NCME. 2014. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Assessment Reform Group. 2002. Retrieved from https://www.researchgate.net/publication/271849158_Assessment_for_Learning_10_Principles_Research-based_principles_to_guide_classroom_practice_Assessment_for_Learning.

Bachman, Lyle. 1990. *Fundamental Considerations in Language Testing and Assessment.* Oxford: Oxford University Press.

Biggs, John. 1996. "Enhancing Teaching through Constructive Alignment." *Higher Education* 32: 347-364. Kluwer Academic Publishers. Retrieved from https://pdfs.semanticscholar.org/1afb/76ca34395777576d5e2c17dcb96c12c77a65.pdf.

Canale, Michael, and Merrill Swain. 1980. "Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing." *Applied Linguistics* 1: 1-47.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge: Cambridge University Press.

—. 2018. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors.* Retrieved from http://www.coe.int/lang-cefr.

Erickson, Gudrun. 2016. "Bedömningens dubbla funktion–för lärande och likvärdighet" ["The Dual Function of Assessment–for Learning and Equity"]. *Communicare–et fagdidaktisk tidskrift fra Fremmedspråksenteret*: 4-9. Retrieved from https://www.fremmedspraksenteret.no/communicare.

Erickson, Gudrun, and Jan-Eric Gustafsson. 2005. *Some European Students' and Teachers' Views on Language Testing and Assessment–a Report on a Questionnaire Survey.* Retrieved from http://www.ealta.eu.org/documents/resources/enlta%20activity%204%20report.pdf.

Erickson, Gudrun, and Åsa Sebestyén. 2018. *Hur kan bedömning främja språklärande och muntlig språkutveckling?* [How can Assessment Enhance Language Learning and the Development of Oral Language Proficiency?] Article, part of "The Language Leap", a web-based professional development program for teachers of modern languages (module: Oral Language Proficiency; Part 8). The Swedish National Agency for Education. Retrieved from https://www.skolverket.se/skolutveckling/kompetensutveckling/sprakspranget---kompetensutveckling-for--larare-i-moderna-sprak%20-%20h-SpraksprangetpaLarportalen.

Erickson, Gudrun, and Lisbeth Åberg-Bengtsson. 2012. "A Collaborative Approach to National Test Development". In *Collaboration in Language Testing and Assessment,* edited by Dina Tsagari, and Ildikó Csépes, 93-108. Frankfurt: Peter Lang.

European Commission. 2012. *First European Survey on Language Competences (Executive Summary).* Retrieved from http://www.surveylang.org/media/ExecutivesummaryoftheESLC_210612.pdf.

Fulcher, Glenn. 2012. "Assessment Literacy for the Language Classroom." *Language Assessment Quarterly* 9(2): 113-132. DOI: 10.1080/15434303.2011.642041.

Gardner, John (ed.). 2012. *Assessment and Learning* (2nd edition). London: Sage.

Green, Anthony. 2013. "Washback in Language Assessment." *International Journal of English Studies* 13(2): 39-51. https://doi.org/10.6018/ijes.13.2.185891.

Hasselgreen, Angela, Cecilie Carlsen, and Hildegunn Helness. 2004. *European Survey of Language Testing and Assessment Needs.* Report: Part One–General Findings. Retrieved from http://www.ealta.eu.org/resources.htm.

Huhta, Ari, Tuija Hirvelä, and Jayanti Banerjee. 2005. *European Survey of Language Testing and Assessment Needs.* Report: Part Two–Regional Findings. Retrieved from http://www.ealta.eu.org/resources.htm.

Hymes, Dell. 1972. "On Communicative Competence." In *Sociolinguistics,* edited by John Pride, and Janet Holmes, 269-293. London: Penguin.

Inbar-Lourie, Ofra. 2008. "Constructing a Language Assessment Knowledge Base: A Focus on Language Assessment Courses." *Language Testing* 25(3): 328-402.

Kane, Michael. 2006. “Validation”. In *Educational Measurement* (4th edition), edited by Robert L. Brennan,17-64. Westport CT: American Council on Education/Praeger Publishers.

Little, David. 2009. *The European Language Portfolio: where Pedagogy and Assessment Meet.* Retrieved from https://www.coe.int/en/web/portfolio/elp-related-publications.

Little, David, and Gudrun Erickson. 2015. “Learner Identity, Learner Agency, and the Assessment of Language Proficiency: Some Reflections Prompted by the Common European Framework of Reference for Languages.” *Annual Review of Applied Linguistics*, 120-139.

Messick, Samuel. 1989. “Validity.” In *Educational Measurement* (3rd edition), edited by Robert Linn, 3-103. New York: American Council on Education/Macmillan.

—. (1996). “Validity and Washback in Language Testing.” *Language Testing* 13: 241-256.

Moss, Pamela, Brian Girard, and Laura Haniford. 2006. “Validity in Educational Assessment.” *Review of Research in Education* 30(1): 109-162. London: Sage Publications.

Sadler, Royce. 1989. “Formative Assessment and the Design of Instructional Systems.” *Instructional Scienc*e 18: 119-144.

Shohamy, Elana. 2001. *The Power of Tests. A Critical Perspective on the Uses of Language Tests.* Harlow, Essex: Pearson Education Limited.

Stiggins, Rick. 1995. “Assessment Literacy for the 21st Century.” *Phi Delta Kappan* 77(3): 238-245.

Stobart, Gordon. 2012. “Validity in Formative Assessment.” In *Assessment and Learning*, edited by John Gardner, 233-242. London: Sage.

Takala, Sauli, Gudrun Erickson, Neus Figueras, and Jan-Eric Gustafsson, 2016. “Future Prospects and Challenges in Language Assessments.” In *Contemporary Second Language Assessment, Contemporary Applied Linguistics*, edited by Dina Tsagari, and Jayante Banerjee, 299-315. London: Bloomsbury Academic.

Torrance, Harry. 2007. “Assessment as Learning? How the Use of Explicit Learning Objectives, Assessment Criteria and Feedback in Post-secondary Education and Training can Come to Dominate Learning.” *Assessment in Education* 14(3): 281-294.

Tsagari, Dina, and Karin Vogt. 2017. “Assessment Literacy of Foreign Language Teachers around Europe: Research, Challenges and Future Prospects.” *Papers in Language Testing and Assessment* 6(1): 41-63.

Weir, Cyril. 2005. *Language Testing and Validation: An Evidence-based Approach*. Basingstoke: Palgrave Macmillan.

Part II:

Understanding Language Assessment Literacy

CHAPTER THREE

AN EMPIRICAL STUDY ON ROMANCE LANGUAGE TEACHERS' SUBJECTIVE THEORIES REGARDING ASSESSMENT PURPOSES AND GOOD TESTS

ANASTASIA DRACKERT, CARMEN KONZETT-FIRTH, WOLFGANG STADLER AND JUDITH VISSER

This chapter presents selected results of an empirical study investigating language assessment practices and the language assessment training of in-service secondary school teachers of French, Italian, and Spanish in Austria and Germany. The focus in this paper is on Romance language teachers' subjective theories regarding classroom-based assessment. The data were collected using an online questionnaire developed by Drackert and Stadler (2017) and adapted for Romance languages. The survey results not only give detailed information about teachers' (*N*=613) reported practices but also inform about their subjective theories on the purposes of assessment and on the criteria of "good" assessment. In addition, the teachers' answers give insights into how important they consider different aspects of LAL. The results lead to a number of implications for research and for teacher education, calling for further studies into the language assessment literacy skills of different teaching contexts and target languages and for interactive teaching methods that involve the embedding of (trainee) teachers' subjective theories in the educational process.

1. Introduction

This chapter presents the results of a survey conducted among "regular"[1] in-service secondary school teachers of Romance languages

[1] By "regular" we mean secondary school teachers who have undergone university training and education and who teach Romance languages at secondary schools.

(French, Italian, and Spanish) in Austria and Germany, elucidating their language assessment literacy (LAL) with regard to assessment purposes, concepts of a "good" test and perceptions of expertise needed at work. Until now, most research both in Europe and overseas has focused on English as a foreign or second language—see, for example, Vogt et al. (2018), dealing with the language assessment literacy of teachers in Cyprus, Greece, Germany, and Hungary; Rodriguez and Zabala (2016), highlighting the situation in Spain, and Kvasova and Kavytska (2014) in Ukrainian universities, while Lam (2015), Baker and Riches (2017) and Zhang and Yan (2018), among others, depict the circumstances in Hong Kong, Haiti, and China. Russian as a foreign language was included in a recent study by Drackert and Stadler (2017, 2018), pinpointing the assessment practices of teachers in Austria, Germany, Switzerland, and South Tyrol (Italy). Their survey was the first to focus entirely on secondary school teachers, in contrast to previous studies which did not separate school teachers from other stakeholders such as university language instructors, teacher trainers or testing experts (for instance, the study by Hasselgreen et al. 2004). Drackert and Stadler's (2017, 2018) results show that both the language taught as well as contextual factors can influence teachers' subjective theories and assessment practices. This seems to be related to terminological and conceptual divergences regarding language assessment between different language (teaching) cultures. It was therefore an obvious next step to investigate the situation in other languages taught at secondary schools. The present survey of teachers of Romance languages also focuses solely on school teachers and complements previous studies on European teachers (e.g., Vogt and Tsagari 2014; Tsagari and Vogt 2017), filling a gap by studying teachers' LAL in French, Italian, and Spanish as a second or third foreign language.

Harding and Kremmel (2016, 424) state that "[a] teacher in Hong Kong may need a very different type of training in assessment from a teacher working in Austria". This paper addresses this issue by emphasizing three research questions:

(1) Why do teachers of Romance languages in Austria and Germany test and assess?
(2) What are their beliefs about a "good" test?
(3) What are their conceptions of skills and abilities needed "to do the job"?

65% of them have additional experience as they collaborate either on the Austrian *Matura* or the German *Abitur* (both school-leaving exams).

2. Language Assessment Literacy: definition and previous research

Although in the research literature LAL has been considered to be "an integral part of teacher professionalism" for quite some time (Xu and Brown 2016, 149) or, as Popham (2009, 4) suggested, "a 'sine qua non' for today's competent educator", there is still little agreement on what assessment literacy might actually comprise. Scholars like Leighton et al. (2010, 9) even claim that assessment literacy is still less widespread than one would assume.

Fulcher's definition of assessment literacy is considered to be the most "comprehensive", most "expanded" and most "detailed" (Inbar-Lourie 2013, 305; Harding and Kremmel 2016, 414; Tsagari and Vogt 2017, 43), as it comprises large-scale assessment and classroom-based tests. As Austrian and German foreign language teachers have a "highly-complex dual role of teacher and assessor" (Inbar-Lourie 2013, 302), the repertoire of components listed by Fulcher seems to be applicable to their necessary understanding of classroom practices as well as to the comprehension of high-stakes standardized testing such as the *Matura* or *Abitur*. Among this repertoire, we find *knowledge, skills and abilities* (regarding the design, development and evaluation of tests), *familiarity* (with test processes), and *awareness* (of test principles, concepts, contextualization and impact) (cf. Fulcher 2012, 125). What we miss in Fulcher's list is the selection of suitable test material and choosing appropriate assessment methods and grading procedures (cf. Harding and Kremmel 2016, 416) as well as an understanding of and a willingness for working in a team in testing and assessing matters. In order to facilitate classroom-based assessment, Cumming (2009, 516) suggests three issues in need of further development: "increasing professional knowledge and abilities; connecting classroom assessment to relevant policies and utilizing teachers' assessment to promote students' learning".

Xu and Brown's framework TALiP[2] (2016) also consists of three levels: knowledge and principles (*what, why and how*); skills and personal understanding (*how assessment "should be"*) and awareness of processes and one's own identity as an assessor (ibid., 159). If knowledge, principles, skills and awareness are entirely dominated or overruled by teachers' individual perceptions, their subjective beliefs, assumptions, and presumed expertise accumulated through "mere work experience" (Arras 2009, 170), teacher-based assessment runs the risk of being biased,

[2] Teacher Assessment Literacy in Practice.

unreliable and invalid. Yet invalid inferences of test results "almost always contribute to poor classroom decisions" (Popham 2001, 106), and this is something teachers need to be aware of and reflect on. Teachers should at least be conscious of their subjective theories and be capable of verifying them on the basis of valid arguments.

Studies by Hasselgreen et al. (2004), Huhta, Hirvalä and Banerjee (2005), Tsagari (2011), Vogt and Tsagari (2014), Tsagari and Vogt (2017) and Vogt (2018) tried to identify the knowledge and skills in language testing and assessment that teachers acquire during their pre-service training at university and their actual needs for further in-service training courses and workshops. Similar research questions were addressed by Drackert and Stadler (2017, 2018) with teachers of Russian as a foreign language. One of the differences between the participants in the three studies by Vogt (2010), Vogt and Tsagari (2014) and Tsagari and Vogt (2017) and that of Drackert and Stadler (2017, 2018) is that (a certain number of) teachers of the latter survey do have experience with standardized tests in contrast to the former groups, and one might assume that this helps them evaluate classroom tests more critically. Furthermore, the studies disagree in some of their conclusions: while Vogt (2010) states that EFL teachers show no interest in further training in assessment issues related to classroom work and grading, Russian teachers do; however, the latter show hardly any interest in alternative methods of assessment such as self- or peer-evaluation.

Since teacher training and assessment training should be tailored to a particular group and address the needs and subjective theories of this particular group, the aim of this survey is to disclose teachers' subjective theories on assessment purposes, "good" tests and the job expertise of Romance language teachers in Austria and Germany—a group that has not been addressed in the LAL research so far. The results may help us to optimize teacher education in the future, be it pre-service or in-service, since the strength of teachers to reflect on their actions and decisions is an important ability for lifelong learning or at least during their professional career (cf. Arras 2009, 178).

3. Romance languages in secondary schools in German-speaking countries

Austria

In Austria, Romance languages are typically learned as a second foreign language, after English. The three languages French, Italian, and

Spanish are offered across all federal states, with French being the most popular, Italian coming second and Spanish catching up (Dalton-Puffer, Faistauer and Vetter 2011, 183). The relatively high numbers of students learning Italian (more than 10% of the pupils in Austria vs. only 2.6% in Europe on average, EACEA 2017) are related to the close historical and economic ties of Italy as a neighboring country (cf. Kral and Wochele 2006). Most students start learning a Romance language in upper secondary school, whereas relatively few do so in primary or lower secondary school (EACEA 2017). The numbers are the same for academically and vocationally oriented schools. This is relevant because the latter represent an important part of the Austrian school system, as over 70% of all upper secondary students attend this school type (EACEA 2017).

Germany

In Germany, French is part of the curriculum in all federal states and it can be learned from primary to secondary school (Minuth 2016, 508). Despite the growing importance of Spanish, it is still a minor second foreign language with about 1.5 million pupils in 2017/18 (Destatis). Among the Romance languages, Italian is of even less significance (about 51,000 pupils in 2017/18; Destatis). It is usually learned as a third or even fourth foreign language. Additional lessons for heritage speakers ("Herkunftssprachlicher Ergänzungsunterricht") are still of some importance (Reimann 2016, 513). Spanish is often called a "booming" foreign language, as it doubled the number of pupils between 2003 and 2013 (Bär 2016, 554), with some 575,000 learners in 2017/18. However, the importance and the position of the Romance languages, French, Italian, and Spanish depend considerably on the political situation in the federal state in question: in most states, French is the second foreign language after English, whereas in Hamburg and Bremen Spanish occupies this position (Destatis).

4. The study

4.1 Participants

The data for the following study come from 613 Romance language teachers, 501 of which completed the whole questionnaire. Out of the 613 participants 504 work in Germany and 109 in Austria; 358 of them teach French, 194 Spanish and 61 Italian. Most of them teach at academically oriented secondary or high schools (*Gymnasium*) (*N*=389). 64.8% of the respondents indicated that they were involved in the preparation, administration or marking of final exams. Their teaching experience ranges from one to 30 years. About a third of the teachers (31.8%) claim that they spend between 20 and 30% of their time on different assessment activities.

4.2 Data Collection and analysis

A three-part online questionnaire[3] developed by Drackert and Stadler (2017, 2018) was adapted for Romance language teachers and used for data collection. Part 1 includes 12 questions on the teachers' professional background. The 13 questions in Part 2 deal with different aspects of teachers' assessment practices and subjective theories with a focus on classroom assessment: they include questions about guidelines and criteria for assessment, the functions of assessment in language classes, the frequency of assessing different skills and the frequency of using different task types and test formats. Part 3 consists of three questions concerning Romance language teachers' wishes and needs regarding in-service training of assessment procedures.

In order to answer our research questions, four questions from the questionnaire were selected for the analysis, namely Q14—Why (for which purposes) do you use exams, tests and evaluations in your French/Italian/Spanish language classes?; Q16—A "good" test: What does that mean to you?; Q19—Please rate the following objectives for written/oral examinations according to their importance on a scale from "not important at all" (=1) to "very important" (=6); and Q21—Please rate the following competences for your job as a French/Italian/Spanish teacher on a scale from "not important at all" (=1) to "very important" (=6).

Thus, two of the four questions are in the closed format and two are in the open-ended format. The open-ended questions allow teachers to introduce new content and express their opinions beyond the answer options provided. Furthermore, the answers to the open-ended questions enable us to better understand teachers' subjective theories on assessment purposes and "good" tests.

To group teachers' free answers, we used Drackert and Stadler's (2017) categorization from their previous survey as a starting point. Each individual author analyzed the answers to the open-ended questions, then a common categorization was agreed upon, so the labels of the categories

---

[3] The questionnaire can be downloaded from https://tinyurl.com/y6qxskzv.

stem from the authors. For example, answers relating to the purposes of assessment (Q14) such as "*in order 'to make' students go through workbooks and textbooks at home*"; "*with the help of frequent tests, pupils are made to study on a regular basis*"; and *"to influence pupils' continuous participation and vocabulary learning"* were subsumed under the assessment function "discipline and activate learners". The answers concerning teachers' views on what makes a test a "good" one (Q16), for instance, *"consists of difficult and easy tasks"; "is of increasing complexity, so that weak students can finish earlier";* and *"both weak and strong students can show their abilities"* were subsumed under the category "includes tasks for weak and strong students."[4]

The survey was developed in Qualtrics and distributed primarily through regional and national associations of teachers of French, Italian, and Spanish in the two countries. The data collection took place in the fall and winter months of 2017/18.

4.3 Results

The first two questions from the questionnaire provide answers to **Research Question 1** (Why do teachers of Romance languages test and assess?), focusing on teachers' objectives in oral and written exams, tests, and evaluations. Table 1 presents an overview of the distribution of 1,608 answers given in percentages across the identified categories for the whole sample (total), the individual languages (FR, SP, and IT) and the two countries, i.e. Germany (GER) and Austria (A). The numbers in the table refer to the percentage of answers falling into a certain category and not to the percentage of teachers submitting an answer.

Table 1: Purposes of exams, tests, and evaluations in % (*N*=552, 1,608 answers)

Purposes of exams, tests, and evaluations	Total	FR	SP	IT	GER	A
1. to assess the achievement of learning objectives	15.2	16.1	13.3	16.3	16.0	11.9
2. to give grades	13.9	13.4	13.9	16.9	13.2	17.3
3. to receive feedback on one's own teaching	12.6	11.9	14.1	11.4	13.4	8.8

[4] Further examples of teachers' free answers (in German) according to all categories can be found under https://tinyurl.com/y3ohlghj.

4. to assess learners' knowledge (primarily grammar and vocabulary)	11.3	12.4	10.0	8.4	11.9	8.2
5. to discipline and activate the learners	7.3	6.1	9.2	8.4	7.5	6.5
6. to repeat and deepen skills and knowledge acquired by learners	6.8	6.3	6.3	10.8	5.5	12.6
7. for learners to self-assess themselves	6.7	7.0	6.7	5.4	6.8	6.1
8. to assess particular competences	6.7	6.5	6.9	7.2	6.6	7.1
9. to increase motivation	5.6	6.5	4.3	4.2	5.8	4.8
10. because it is required in the curriculum	4.9	4.4	6.5	2.4	4.3	7.1
11. to give feedback on learners' individual progress	4.1	4.5	4.1	1.8	4.4	2.7
12. to assess learners' individual progress	3.2	3.2	2.9	3.6	2.9	4.4
13. to prepare for the school-leaving exam	1.3	1.0	1.4	3.0	1.1	2.4
14. for the school's internal evaluation	0.4	0.5	0.4	0.0	0.5	0.0

As can be seen from Table 1, fourteen categories of assessment purposes are present in teachers' answers. The three most frequently reported purposes of classroom-based assessment are 1) to assess the achievement of learning objectives (15.2%), 2) to give grades (13.9%), and 3) to receive feedback on one's own teaching (12.6%).

Three of the categories (1, 4 and 8) are fairly general. Category 1 seems more of a pretext than a purpose, as it remains unclear how language teachers use the results of the tests. The same generalization applies to Categories 4 and 8. Category 4 combines teachers' answers such as assessing grammar, vocabulary, orthography, and pronunciation, in other words, answers that represent areas of knowledge, whereas the answers included in Category 8 focus on assessing particular skills such as speaking or reading, and both focus on the *what* and not on the *why*. All three categories make up 33.2% of the teachers' responses.

Three further groups of answers frequently mentioned by the teachers concern the use of assessment for formative purposes, namely for learners to self-assess themselves (Category 7), to give feedback on learners' individual progress (Category 11), and to assess learners' individual progress (Category 12). These three purposes, if combined, make up 14% of all answers and thus seem to be of equal importance to the teachers as the purpose of giving grades on the basis of assessments (13.9%).

Two further functions of assessment play an important role in teachers' opinions: using assessment to activate (Category 5) and motivate learners

(Category 9). Both categories make up 12.9% of the answers and are, in sum, as essential as receiving feedback to one's own teaching with the help of assessment (Category 3). A total of 6.8% of the answers indicates that teachers also use assessment so that students practice their skills and deepen their knowledge.

4.9% of the answers amount to the fact that assessment is required in the curriculum (Category 10), in other words, testing and assessing are not necessarily the teachers' free choice. Using assessment to prepare learners for the school-leaving exams (Category 13) is mentioned in only 1.03% of all answers. This is surprising, particularly for the Austrian context, as the introduction of the national standardized school-leaving exam has demonstrably led to increased teaching-to-the-test in language classrooms (cf. Frötscher 2016).

Additional cross-tabulation analyses enabled us to shed light on the differences in teachers' perceptions across the two countries. In Austria, the most common purposes of assessment seem to be 1) giving grades (17.3%), 2) strengthening (deepening) acquired skills and knowledge (12.6%), and 3) assessing the achievement of learning objectives (11.9%), whereas in Germany the order is quite different. Furthermore, the use of exams, tests, and evaluations because it is required in the curriculum (7.1%) seems to be a greater issue in Austria than in Germany, where this category makes up only 4.3% of the answers. German teachers in turn consider it more important to assess students' knowledge of grammar and vocabulary (11.9%), whereas only 8.2% of the answers of Austrian teachers fall into this category.

The differences in teachers' perceptions of the purposes of assessment also seem to depend on the language(s) they teach. For instance, Spanish teachers consider receiving feedback to their teaching as the most important function of assessment (14.1% vs. 11.9% for French and 11.4% for Italian). French teachers, however, mention the assessment of linguistic knowledge (primarily of grammar and vocabulary) more often than Spanish and Italian teachers (12.4% vs. 10% vs. 8.4%). At the same time, French teachers report using assessment as a means of student self-assessment and increasing students' motivation more often than their partners. And Italian teachers consider the use of assessment so that students practice their skills and deepen their knowledge to be more important than their Spanish or French colleagues (10.8% vs. 6.3%). Also, using assessments to give grades is a more frequent answer with Italian teachers than with their French or Spanish partners (16.9 % vs. 13.4% and 13.9%).

Q19 addresses the same issue, i.e. the functions of tests and assessments, but in the closed format. Table 2 shows the mean values and standard deviations (in brackets) of teachers' agreement with certain purposes of assessment, presented for the whole sample, individual languages and two countries.

Table 2: Teachers' ranking of the purposes for implementing exams, tests, and evaluations (*N*=525)

	I assess learner performances	Total	FR	SP	IT	GER	A
1.	to give learners individual feedback on their strengths and weaknesses	5.09 (1.43)	5.06 (1.46)	5.07 (1.43)	5.36 (1.23)	5.10 (1.41)	5.05 (1.52)
2.	to check whether learners have achieved the learning objectives	5.02 (1.40)	4.99 (1.46)	5.07 (1.30)	5.04 (1.37)	5.00 (1.39)	5.10 (1.48)
3.	to be able to address mistakes systematically	4.68 (1.38)	4.73 (1.37)	4.54 (1.40)	4.79 (1.38)	4.65 (1.39)	4.82 (1.33)
4.	so that learners reflect on their own learning process and progress	4.68 (1.41)	4.70 (1.42)	4.60 (1.43)	4.85 (1.31)	4.67 (1.38)	4.75 (1.54)
5.	to give parents specific information about their children's performance	3.86 (1.48)	4.06 (1.46)	3.64 (1.49)	3.42 (1.41)	3.92 (1.48)	3.61 (1.47)
6.	because progress checks have to be conducted continuously	3.81 (1.56)	3.92 (1.57)	3.58 (1.59)	3.89 (1.42)	3.86 (1.54)	3.60 (1.66)
7.	because this way my students study harder and more	3.78 (1.39)	3.91 (1.36)	3.63 (1.46)	3.53 (1.25)	3.83 (1.38)	3.56 (1.41)
8.	because my students expect me to	2.79 (1.53)	2.87 (1.55)	2.74 (1.53)	2.47 (1.39)	2.84 (1.56)	2.55 (1.38)
9.	to discipline the group	2.71 (1.47)	2.78 (1.47)	2.72 (1.46)	2.23 (1.38)	2.76 (1.44)	2.46 (1.55)

Table 2 shows that Romance language teachers consider the following four purposes of assessment as the most important: 1) to give learners individual feedback on their strengths and weaknesses, 2) to check whether learners have achieved the learning objectives, 3) to address mistakes systematically, and 4) to give learners the opportunity to reflect on their learning process and progress. Of these four functions, the third one—to address mistakes systematically—was not explicitly mentioned by the teachers in their answers to the open-ended question. Bringing discipline into the classroom (Category 9) was considered the least important purpose of assessment, although this category ranks fifth in the open-ended question.

If the answers to this question are compared across languages and countries, only a few differences emerge. For example, teachers of Spanish and Italian attribute more importance to assessing learner performances so that students reflect on their learning process and progress (Category 4), while French teachers put more weight on the systematic addressing of mistakes (Category 3). Furthermore, teachers of Spanish have the highest belief that tests help their students learn harder and study more (Category 7); they consider this a more important reason for testing than the institutionally predetermined need to do so. Finally, to Romance language teachers in Germany, checking whether their students have achieved the learning objectives (Category 2) is the most important function of assessment; to teachers in Austria, however, the most relevant function of tests is to give students individual feedback on their strengths and weaknesses (Category 1).

Another key area of beliefs influencing the way tests are developed or tasks are selected is teachers' perceptions of a "good" test. Table 3 presents an overview of the distribution of 881 answers to **Research Question 2** given in percentages across the identified criteria for the whole sample, the individual languages and the two countries separately.

Table 3: Teachers' subjective theories on the criteria of a "good" test in % (*N*=524, 881 answers)

	Criterion	Total	FR	SP	IT	GER	A
1.	construct validity	20.7	19.4	23.1	19.5	20.4	21.8
2.	assesses different competences	10.8	10.3	11.9	9.8	11.4	7.7
3.	corresponds to lesson content and proficiency level of the learner group	10.7	9.9	12.2	9.8	10.9	9.6
4.	test in which students do well	10.0	12.9	5.8	7.3	10.8	6.4
5.	is transparent	9.8	11.1	6.8	12.2	9.4	11.5

6.	includes tasks for weak and strong students	7.2	7.1	7.1	7.3	7.3	6.4
7.	uses different test formats	7.0	7.1	6.4	8.5	7.2	6.4
8.	is feasible	6.8	6.5	7.1	7.3	6.2	9.6
9.	gives feedback on one's own teaching	6.4	5.2	7.5	9.8	5.5	10.3
10.	provides students with helpful feedback	4.2	4.2	5.1	1.2	4.8	1.3
11.	is objective	2.8	2.2	3.4	4.9	2.2	5.8
12.	is exciting and interesting for students	2.2	1.8	3.1	1.2	2.3	1.3
13.	is designed appropriately	1.2	1.8	0.3	1.2	1.4	0.6
14.	other	0.3	0.4	0.3	0.0	0.1	1.3

As can be seen from Table 3, the most important criterion of "good" assessment mentioned by the teachers is construct validity, which comprises 20.7% of the answers. Four further criteria of a "good" test reported in approximately 10% of the responses are: assessing different competences (Category 2), the fact that the test corresponds to the content of the lesson and the proficiency level of the learners (Category 3), a test in which students do well (Category 4) as well as transparency (Category 5). While about 10% of the answers reveal that the question was understood differently than intended by the authors (i.e. teachers reply that a "good" test is one in which the students do well = Category 4), 7.2% of answers mention the differentiating criterion as a characteristic of a "good" test, namely that the test includes tasks or items both for the weak and the strong students. The use of different test formats and the principle of practicality rank as the sixth and seventh criteria of a "good" test.

According to the teachers, a "good" test should also fulfill the useful functions mentioned in teachers' answers to the previous questions, such as giving feedback to one's own teaching (Category 9) and providing students with helpful feedback (Category 10), which together make up approximately 10% of all answers. Objectivity (Category 11), a test being interesting and exciting for students (Category 12) and an appropriate test design (Category 13) do not seem to be central criteria of a "good" test, since these criteria are mentioned in only one to three per cent of all the answers.

Some of the criteria play a different role depending on which language is taught and in which country. While teachers of Spanish mention construct (Category 1) and content validity as well as the proficiency level of their learners (Category 3) more frequently than their colleagues, teachers of French and Italian list transparency (Category 5) more often. At the same time, teachers of Italian consider the use of different test formats (Category 7), useful feedback to their teaching (Category 9) and objectivity (Category 11) as more important criteria of a "good" test than their partners teaching French and Spanish.

With regard to the differences between the countries, the data also disclose that Romance language teachers in Germany consider the assessing of different competences (Category 2) as more relevant, whereas Romance language teachers in Austria mention transparency (Category 5), feasibility (Category 8), and feedback on their own teaching (Category 9) more often than their German partners.

Finally, we answer **Research Question 3** by presenting a ranking of assessment related competences (see Table 4) that teachers regard as crucial in their job.

Table 4: Teachers' ranking of LAL components needed for the job (*N*=524, 881 answers)

	being able to	Total	FR	SP	IT	GER	A
1.	make assessment transparent and comprehensible	5.42 (1.10)	5.48 (1.03)	5.29 (1.21)	5.57 (0.89)	5.40 (1.08)	5.56 (1.09)
2.	assess students in a fair way	5.40 (1.09)	5.43 (1.04)	5.34 (1.18)	5.57 (0.91)	5.39 (1.08)	5.51 (1.05)
3.	give students feedback on the basis of test scores	5.08 (1.14)	5.09 (1.12)	5.03 (1.18)	5.26 (0.98)	5.05 (1.14)	5.25 (1.07)
4.	evaluate the impact of instruction	5.06 (1.12)	5.05 (1.09)	5.07 (1.19)	5.19 (0.94)	5.06 (1.10)	5.10 (1.13)
5.	know students' strengths and weaknesses	5.06 (1.11)	5.06 (1.09)	5.06 (1.18)	5.15 (1.01)	5.07 (1.10)	5.07 (1.05)
6.	correct learners' mistakes	4.96 (1.16)	4.99 (1.11)	4.88 (1.23)	5.08 (1.14)	4.96 (1.13)	5.00 (1.26)
7.	give grades	4.80 (1.31)	4.81 (1.32)	4.80 (1.30)	4.83 (1.19)	4.82 (1.31)	4.76 (1.24)
8.	create test tasks	4.68 (1.36)	4.56 (1.32)	4.79 (1.43)	5.04 (1.16)	4.67 (1.35)	4.75 (1.35)
9.	evaluate the impact of the assessment procedures used	4.50 (1.31)	4.50 (1.30)	4.48 (1.35)	4.70 (1.17)	4.50 (1.33)	4.58 (1.21)

10.	carry out informal assessment methods (self- or peer assessment, portfolios, learning diaries, etc.)	3.62 (1.37)	3.60 (1.35)	3.68 (1.39)	3.64 (1.46)	3.64 (1.37)	3.59 (1.38)
11.	use the CEFR and the ELP	3.53 (1.42)	3.54 (1.44)	3.47 (1.39)	3.70 (1.32)	3.41 (1.37)	4.07 (1.45)
12.	detect every mistake	3.38 (1.40)	3.36 (1.43)	3.55 (1.33)	3.00 (1.30)	3.45 (1.37)	3.11 (1.49)

As can be seen from Table 4, two of the competences related to assessment activities that teachers rank highest are 1) to make assessment transparent and comprehensible (*M*=5.42), and 2) to assess students in a fair way (*M*=5.40). Similarly, also important to them (*M*=4.08 and *M*=4.06) are competences such as giving students feedback on the basis of test scores (Category 3) and evaluating the impact of instruction and knowing students' strengths and weaknesses (Category 4). The competences that the teachers consider least important are carrying out informal assessment methods (Category 10), using the *Common European Framework of Reference for Languages* and the *European Language Portfolio* (Category 11) and to detect every single mistake (Category 12).

With regard to these questions, the data exhibit only two differences between the teachers depending on language and country: overall, teachers of Italian rank competences higher than teachers of French and Spanish, with two exceptions (being able to carry out informal assessment methods (Category 11) and detecting every single mistake (Category 12)). The greatest divergences in the ratings between the teachers of different languages can be observed in the ability to create test tasks, which the teachers of Italian consider to be the most important (*M*=5.04), followed by the teachers of Spanish (*M*=4.79) and the teachers of French (*M*=4.56).

The competences ranked higher by Austrian teachers than by their German colleagues were making assessment transparent and comprehensible (Category 1), giving students feedback on the basis of test scores (Category 3) and using the *CEFR* and the *ELP* (Category 11), whereas the ability to give grades (Category 3) was ranked higher by German teachers.

5. Discussion and Conclusion

The present survey of school teachers of Romance languages aims at complementing previous studies on classroom-based assessment (e.g. Vogt and Tsagari 2014; Tsagari and Vogt 2017) and at filling a gap by studying teachers' LAL in French, Italian, and Spanish as a second or third foreign language. The high turnout allows us not only to give detailed information about teachers' preconcept(ualizat)ions, but also to inform about their "personalized knowledge" (Scarino 2017, 36) or subjective theories (cf. Groeben et al. 1988) on the purpose of assessment practices and criteria of good assessment. The teachers' answers reveal insights into their views about the importance of different components of language assessment literacy. Thus, the results of the study have extended our knowledge in the area of language assessment literacy in general and they enable us to make concrete suggestions for the pre- and in-service training of teachers of Romance languages in particular.

5.1 Implications for LAL research

As regards Research Question 1, namely why teachers of Romance languages test and assess, we were impressed by the variety of test purposes listed by the teachers. This remarkable spectrum shows that teachers of Romance languages do not simply test for testing's sake or because it is required by the authorities. Yet, in quite a considerable number of responses (13.2% of mentions in Austria, 17.3% in Germany) teachers state that they test to give grades—a fact that represents an institutional demand. But when these same teachers are given a range of test purposes to judge and rank (cf. Table 2), bureaucratic reasons figure very low in importance. Instead, the purposes that centre around their students' progress, diagnosis and feedback come out on top. Since these formative functions of assessment seem vital to teachers, they should be a fundamental part of teacher training (see 5.2).

Research Question 1 was dealt with in two questions, a closed ranking format and an open question: Q14—"Why (for which purposes) do you use exams, tests and evaluations in your French/Italian/Spanish language classes?" The open question clearly left considerable room for teachers' interpretation: judging from the data, teachers do not have theory-driven answers ready for this question; instead, they seem to draw from their subjective and practice-driven beliefs about testing and assessment. In their subjective theory, the *what* and the *why* of assessment (cf. Xu and Brown 2016) may well converge, but how this is or is not compatible with evidence-based theory is, of course, another matter, which might be worth exploring. Besides, we must not forget that subjective theories are one thing, while actual practice is another. Teachers' beliefs might align with their practices but this need not always be the case (cf. Aristizábal 2018, 40). There is a need for more research into the convergence or divergence of the two, for instance by following up questionnaire studies such as ours with direct investigations into the concrete, situated practices of teachers' classroom-based assessment.

When we relate our findings to Norris' (2012) description of "purposeful language assessment", approximately 30% of teachers' answers seem to indicate not necessarily a lack of purposes but rather a lack of purposefulness, i.e. the knowledge that can help them distinguish between the "jobs to be accomplished", the role of assessment in these jobs and the nature and use of assessment tools. Testing and assessment seem to take on something like an "all-round function" for Romance language teachers. Consider, for instance, the fact that tests are given to practice skills and deepen knowledge (cf. Table 1, Category 6). At first sight, such multi-tasking is not surprising in a classroom-based context, and it is not unacceptable *per se*; however, if a test or another form of assessment fulfills so many functions, possibly concurrently, one must ask if teachers actually know how to use tests to respond to all these demands.

Research Question 2 was covered by the questionnaire item "What is, in your opinion, a 'good' test?" The answers given reveal further shades of teachers' subjective theories. Construct validity is the most frequently mentioned criterion embedded in teachers' answers, sometimes explicitly described as "the test should test what it is supposed to test". At the same time, similar to the teachers of Russian as a foreign language in Germany and Austria surveyed by Drackert and Stadler (2017, 2018), many Romance language teachers say that a "good" test should somehow correspond to the current level of the learner group. From a test theoretical perspective, this is problematic—can a test be construct-valid if it is at the same time adjusted to the group's average competence level? It is possible that teachers either may not be assessment literate enough to recognize this problem or that their answers are built on their own theories about testing and assessment. These subjective theories seem to suggest that tests need to be very closely connected with the teaching, the class and its individuals. Secondary school teachers—in their dual role as teacher-testers—do indeed have a responsibility for their test takers that goes far beyond delivering the test and its results. In their subjective theories as presented to us in their answers, the test seems to be very much part and parcel of their teaching, including disciplinary and motivational, even acquisition-related factors.

A further similarity to the results of Drackert and Stadler's (2017, 2018) study is the importance of dealing with differentiation in classroom tests: according to many teachers, a "good" test should be manageable both by weak and strong students. This does not only point to a fundamental difference between classroom-based and standardized tests (Dlaska and Krekeler, 2009), but it also highlights the need for different stakeholders to acquire assessment literacy training that is tailored to their individual needs.

The differences in views about language testing and assessment between teachers of different languages (consider, for instance, that Italian teachers see test development or the use of tests to receive feedback to one's own teaching as more important than teachers of Spanish or French) allow us to conclude that the specific teaching context—including the language(s) taught—influences teachers' beliefs and practices (cf. Drackert and Stadler 2017, 2018), and this needs to be accounted for both in research and training. More generally speaking, the results of our survey amount to a plea for more LAL studies within different teaching cultures and different target languages in order to understand more clearly which factors influence teachers' decision-making processes in assessment situations and to fine-tune the language assessment content of teacher education programs.

5.2 Implications for language teacher education

The results of our study highlight the significant role of subjective theories held by Romance language teachers. Subjective theories based primarily on experience (Arras 2009, 170) fill the gap between what we perceive as teachers' lack of knowledge about language testing and assessment and the existing need to test and assess in the classroom on a daily basis.

Most likely, subjective theories will never be completely eradicated from teachers' professional lives but as the findings of our study demonstrate, there is certainly room for research-based theory in order to support teachers' practical choices. Pre-service (and, to a lesser extent, in-service) teacher education plays an important role in this process. The answers teachers give to the questions in our study suggest that many of them have not been sufficiently trained in matters of testing and assessment. Gaps in their language assessment literacy give way to the development and consolidation of subjective theories.

It is vital that education curricula for Romance language teachers stipulate a thorough training in testing and assessment, focusing specifically on the role and challenges of tests and assessment in classroom contexts. While the core testing principles of validity and reliability are, of course, essential in a course on language testing, they must not remain abstract terms but need to be explored in the context of classroom practice. Many of the answers in our study show that in-service teachers of Romance languages find it difficult to see through theoretical key concepts (which they are often familiar with) and their practical implementation for classroom use.

Judging from the results of our study,

1) language testing theory needs to be adapted to suit classroom-based assessment issues first and foremost;
2) there needs to be a (stronger) focus on classroom-based assessment literacy covering aspects such as what tests can "do" in the classroom and what they cannot (or what is better achieved by applying other measures);
3) there also needs to be a clear focus on the realities and complexities of classroom-based assessment, on practical issues such as time constraints and institutional resources, and how to deal with them (e.g. how to deal with feedback and which feedback to give to which stakeholders) and how to deal with heterogeneous groups;
4) teacher training should address subjective theories of a particular target group (e.g. Romance language teachers or sub-groups such as Italian or French or Spanish teachers) and work with the (trainee) teachers to reflect on and verify these theories on the basis of scientific evidence.

6. References

Aristizábal, Frank G. 2018. "A Diagnostic Study on Teachers' Beliefs and Practices in Foreign Language Assessment." In *Íkala, Revista de Lenguaje y Cultura* 23(1):25-44.

Arras, Ulrike. 2009. "Subjektive Theorien als Faktor bei der Beurteilung fremdsprachlicher Kompetenzen." In *Sprachlehrforschung: Theorie und Empirie. Festschrift für Rüdiger Grotjahn*, edited by Annette Berndt, and Karin Kleppin, 169-179. Frankfurt am Main.: Peter Lang Verlag.

Baker, Beverly A., and Caroline Riches. 2017. "The Development of EFL Examinations in Haiti: Collaboration and Language Assessment Literacy Development." *Language Testing* 35(4):557-581.

Bär, Marcus. 2016. "Spanisch". In *Handbuch Fremdsprachenunterricht,* edited by Eva Burwitz-Melzer et al., 553-558. Tübingen: Francke.

Cumming, Alister. 2009. "What Needs to be Developed to Facilitate Classroom-based Assessment?" *TESOL Quarterly* 43(3):515-519.

Dalton-Puffer, Christiane, Renate Faistauer, and Eva Vetter. 2011. "Research on Language Teaching and Learning in Austria (2004–2009)." *Language Teaching* 44(2):181-211.

Destatis Statistisches Bundesamt. 2019. "Schüler/innen mit fremdsprachlichem Unterricht." Accessed February 24, 2019. https://www.destatis.de/DE/ZahlenFakten/GesellschaftStaat/BildungForschungKultur/Schulen/Tabellen/AllgemeinBildendeBeruflicheSchulenFremdsprachUnterricht.html.

Dlaska, Andrea, and Chrisian Krekeler. 2009. *Sprachtests. Leistungsbeurteilungen im Fremdsprachenunterricht evaluieren und verbessern.* Baltmannsweiler: Schneider Verlag Hohengehren.

Drackert, Anastasia, and Wolfgang Stadler. 2018. "Assessment Literacy of Russian as a Foreign Language Teachers." In *Current Trends and Future Perspectives in Russian Studies. Актуальные проблемы и перспективы русистики*, edited by Joan Castellví, Andrey Zainouldinov, Ivan Garcia, and Marc Ruiz-Zorrilla, 963-973. Barcelona: Trialba Ediciones.

—. 2017. "Beurteilungskompetenz von Russischlehrkräften in Deutschland, Österreich, der Schweiz und Südtirol (DACHS): Zwischen Status Quo und aktuellen Bedürfnissen." *Zeitschrift für Fremdsprachenforschung* 1(28):233-258.

Europäische Kommission/EACEA/Eurydice. 2017. *Schlüsselzahlen zum Sprachenlernen an den Schulen in Europa–Ausgabe 2017.* Eurydice-Bericht. Luxemburg: Amt für Veröffentlichungen der Europäischen Union.

Frötscher, Doris. 2016. "A New National Exam: A Case of Washback." In *Contemporary Second Language Assessment. Contemporary Applied Linguistics Volume 4*, edited by Dina Tsagari, and Jayanti Banerjee, 61-81. London/New York: Continuum.

Fulcher, Glenn. 2012. "Assessment Literacy for the Language Classroom." *Language Assessment Quarterly* 9(2): 113-132.

Groeben, Norbert, Diethelm Wahl, Jörg Schlee, and Brigitte Scheele. 1988. *Forschungsprogramm Subjektive Theorien. Eine Einführung in die Psychologie des reflexiven Subjekts*. Tübingen: Francke.

Harding, Luke, and Benjamin Kremmel. 2016. "Teacher Assessment Literacy and Professional Development." In *Handbook of Second Language Assessment. Handbooks of Applied Linguistics, Vol. 12*, edited by Dina Tsagari, and Jayanti Banrejee, 413-428. Boston/Berlin: Mouton De Gruyter.

Hasselgreen, Angela, Cecile Carlsen, and Hildegunn Helness. 2004. "European Survey of Language Testing and Assessment Needs. Part One: General Findings." Accessed February 24, 2019. http://www.ealta.eu.org/documents/resources/survey-report-pt1.pdf.

Huhta, Ari, Tuija Hirvalä, and Jayanti Banerjee. 2005. "European Survey of Language Testing and Assessment Needs. Part Two: Regional Findings." Gothenburg: European Association for Language Testing and Assessment.

Inbar-Lourie, Ofra. 2008. "Constructing a Language Assessment Knowledge Base: A Focus on Language Assessment Courses." *Language Testing* 25(3):385-402.

Kral, Pia, and Holger Wochele. 2006. "Fremdsprachen- und Fehlerbewusstheit von Laien. Empirische Untersuchungen am Institut für Romanische Sprachen der Wirtschaftsuniversität Wien." In *Romanische Sprachen in Europa. Eine Tradition mit Zukunft? Akten zur gleichnamigen Sektion des XXIX. Deutsche Romanistenages an der Universität des Saarlandes (25.–29. September 2005)*, edited by Michael Frings, and Andre Klump, 192-208. Stuttgart: ibidem-Verlag.

Kvasova, Olga, and Tamara Kavytska. 2014. "The Assessment Competence of University Language Teachers: A Ukrainian perspective." *CercleS* 4(1):159-177.

Lam, Ricky. 2015. "Language Assessment Training in Hong Kong: Implications for Language Assessment Literacy." *Language Testing* 32(2):169-197.

Leighton, Jacqueline P., Rebecca J. Gokiert, M. Ken Cor, and Colleen Heffernan. 2010. "Teacher Beliefs about the Cognitive Diagnostic Information of Classroom- versus Large-Scale-Tests: Implications for Assessment Literacy." *Assessment in Education: Principles, Policy & Practice* 17(1):7-21.

Minuth, Christian. 2016. "Französisch." In *Handbuch Fremdsprachenunterricht*, edited by Burwitz-Melzer, Eva et al., 507-512. Tübingen: Francke.

Rodriguez, Cristina, and Julia Zabala. 2016. "Language Assessment Literacy for Test Developers: Towards an Understanding of Test Constructs." Presented in May 2016 at the 13th EALTA Conference, Valencia, Spain. http://www.ealta.eu.org/conference/2016/presentations/Friday/Rodriguez_Zabala.pdf.

Scarino, Angela. 2017. "Language Assessment Literacy as Self-awareness: Understanding the Role of Interpretation in Assessment and in Teacher Learning." *Language Testing* 30(3):309-327.

Tsagari, Dina. 2011. "Investigating the 'Assessment Literacy' of EFL State School Teachers in Greece." In *Classroom-based Language Assessment*, edited by Dina Tsgari, and Ildikó Csepes, 169-190. Frankfurt am Main.: Peter Lang Verlag.

Tsagari, Dina, and Karin Vogt. 2017. "Assessment Literacy of Foreign Language Teachers around Europe: Research, Challenges and Future Prospects." *Papers in Language Testing and Assessment* 6(1):41-63.

Vogt, Karin. 2010. "Assessment Literacy bei Fremdsprachenlehrkräften–Bedarf und Bedürfnisse in Aus- und Weiterbildung." In *Grenzen überschreiten: sprachlich - fachlich - kulturell. Dokumentation zum 23. Kongress für Fremdsprachendidaktik der Deutschen Gesellschaft für Fremdsprachenforschung (DGFF) Leipzig, 30. September–3. Oktober 2009,* edited by Claus Altmayer, Grit Mehlhorn, Christiane Neveling, Karen Schramm, and Norbert Schlüter, 267-277. Baltmannsweiler: Schneider.

—. 2018. "Assessment: The What, Why and How." In *Teaching English as a Foreign Language. An Introduction*, edited by Carola Surkamp, and Barbara Viebrock, 249-266. Stuttgart: Metzler Verlag.

Vogt, Karin, and Dina Tsagari. 2014. "Assessment Literacy of Foreign Language Teachers: Findings of a European Study." *Language Assessment Quarterly* 11(4):374-402.

Xu, Yueting, and Gavin T. Brown. 2016. "Teacher Assessment Literacy in Practice: A Reconceptualization." *Teaching and Teacher Education* 58:149-162.

Zhang, Cong, and Xun Yan. 2018. "Assessment Literacy of Secondary EFL Teachers: Evidence from a Regional EFL Test." *Chinese Journal of Applied Linguistics* 41(1):25-46.

CHAPTER FOUR

GREEK LANGUAGE TEACHERS' CONCEPTIONS ON PURPOSES AND APPROACHES OF ASSESSMENT

GEORGIA SOLOMONIDOU

Assessment is one of the protagonists of the earliest establishment of formal education. Though assessment in its early attempts was largely seen as a measurement instrument, alternative assessment of student achievement has arrived on the scene in the past three decades. In this context, many studies have emerged worldwide investigating students' and teachers' conceptions of assessment approaches and purposes as these will affect their practices. The current study uses a sequential mixed-method design to investigate language teachers' conceptions on the purposes of student assessment in general and the assessment framework specific for the Modern Greek language subject as the first language within the New National Curriculum in Cyprus (NNC) based on widespread ideologies in Language Assessment Literacy (LAL). The methods involved a questionnaire administered to Greek language teachers (N=95) followed by (N=7) individual interviews. The qualitative data from the same sample helped to explain and build upon initial quantitative results. Results are discussed in light of other research showing that the conceptions of teachers are in alignment with the current shift of assessment to be used for enhancing teaching and learning. Teachers tend to agree with the legitimate purposes of assessment whilst they would like to use assessment for learning as also promoted by the NNC.

1. Introduction

Whilst the idea of assessment to support learning is not new, it was not until the late 1980s that the idea that classroom assessment practices could both enhance or even limit student learning started to receive common acceptance (Wiliam 2011). Fundamentals come from the grounds of sociocultural learning theory, and metacognitive and self-regulation theory whilst a number of writers (Heritage 2013; Sadler 1989; Black and Wiliam 1998; Black and Wiliam 2010; Torrance and Pryor 1998) began to dispute more powerfully that the learner had to be assigned a central role in the assessment process. Consequently, during the 1990s the concept of assessment for learning first appeared in the assessment literature and has now become a part of the discourse of assessment (Gardner 2012, cited in Dixon and Haigh 2009). A major change in approaches to assessment is the shift to view assessment not only as a means to an end, to determine measurement and thus certification (as it has generally been in the past) but also as “a tool for learning” (Dochy and McDowell 1997, 279).

Present research into assessment as a tool to support student learning is increasingly focused on how this support is perceived (Pat-El, Tillema, Segers and Vedder 2013). As argued by Butler and McMunn (2006), for improving teachers’ assessment practices we should first observe their beliefs that inspire these practices. As stated by Shepard (2000), for the assessment framework to be reformed the teachers’ conceptions on assessment must change accordingly. Changes in conceptions may even have to occur first before any practical changes occur in teachers’ practices because the way they perceive things is closely related to how they will treat them.

As conceptions on assessment, in Cyprus were largely ignored, it was decided to generate quantitative data for establishing the baseline on conceptions about the purposes of assessment as well as what teachers associate with the term assessment; for placing the island in the scenery of the global discussion. The assessment framework in Cyprus was changing at the time of this study (in 2011-2012) as part of a large-scale curriculum change called the New National Curriculum (NNC) the implementation of which started in 2010. One of the main aims of the New National Curriculum (NNC) strategy is to more fundamentally associate assessment with instruction. It is valuable to mention here that generally the assessment framework in Cyprus was and still is low stakes with no mandatory external exams beyond the end of grade school wide exams which are used to determine whether the student can go on to the next level and the nationwide Year 12 (age 18) exams for entering higher level institutions in Cyprus and Greece.

2. Greek Language Assessment

Modern Greek as the mother language of the country is given much attention by the Cypriot curriculum and the NNC. One main goal expected to be achieved through Modern Greek is for students to develop multiple reading practices which will lead to a critical awareness of the world. Students are also expected to understand the ways in which a specific linguistic and documental choice implies differing social conditions and ideological formations; understanding the way in which different textual articles function in the framework of a school community and other communities for the negotiation or control of specific ideologies. The syllabus also emphasizes the need to broaden the "creativity" of children through their familiarization with multiple writing practices. "Creativity" is seen as the ability of students to express themselves through their written documents and the social relationships with their interlocutor. It is also the route to build ways of resisting modular structures, and it is placed on the highest level of Bloom's revised taxonomy of the cognitive domain. Students are also expected through the NNC, and the Greek language teaching (and consequently assessment) to develop "critical reflection through writing", meaning the way that humans can better understand their surroundings through script (National Curriculum, 3-4, *author's translation*). Throughout the school year, all students should be given the opportunity to produce several species of written documents or oral speech texts addressed to a real listener.

In the philosophy of the Greek language syllabus, which has been reformed as part of the NNC, the relationship between learning and teaching is being viewed as a dynamic process rather than a one-way transmission of knowledge. Children are expected to interact in the classroom, raise questions and actively participate avoiding the notion of being passive learners. Specific actions for the children to undertake in order to be active learners are suggested in the Greek language syllabus. This shows an attempt to involve children not only in the learning process but in the teaching and assessment process too (Askew and Carnell 1998; Biggs and Moore 1993; Watkins et al. 1996, cited in Askew 2000).

The Greek language NNC assessment philosophy, though taught as students' first language here, is based on common ideologies in language assessment literacy (LAL) and broadly here defined as the knowledge, skills but also abilities to design, develop, undertake, evaluate and maintain several assessment strategies in the classroom which are then going to be used for enhancing teaching and learning (Black and Wiliam 1998; Broadfoot 2007; Fulcher 2012; Harlen and James 1997; Malone 2013; Popham 2017; Vogt and Tsagari 2014).

3. Teachers' Conceptions

Teachers' conceptions are crucial since teaching is a personal endeavour; therefore, any changes in teaching or assessment strategies can only be acknowledged when seen as realistic and helpful and can only be welcomed by teachers who are willing to apply them (Dixon and Haigh 2009; Sato et al. 2005; Jones and Moreland 2005). Teachers' conceptions of assessment are important as they affect their practices (Borg 2003; Clark and Peterson 1986; Pajares 1992, cited in Harris and Brown 2009; Kyriakides 1996) which in turn may affect students' outcomes (Harris and Brown 2009). Moreover, teachers' beliefs about students, learning, teaching, and subjects influence their assessment techniques and practices (Brown et al. 2011; Duncan and Noonan 2007). One of the difficulties in researching teachers' conceptions of assessment is that they appear to hold multiple and possibly even contradictory conceptions at the same time. A major factor in this plurality of conceptions is that assessment itself serves multiple purposes, which may be complementary or contradictory (Brown et al. 2011).

The impact of Gavin Brown's research is so wide in the research of this field that the literature available could be divided into studies using the teachers' conceptions of assessment inventory (TCoA) developed by Brown since 2002 and studies which are not using this instrument. Although these studies may overlap, I suggest they take different pathways in investigating teachers' conceptions. Thus, studies using the TCoA imply that there are some universal beliefs about assessment purposes that may reflect the shared low-stakes policy environment of assessment. The teachers' conceptions of assessment (TCoA) inventory (Brown 2002, 2004, 2006, 2008) is a self-reporting multidimensional survey instrument designed to elicit teachers' level of agreement with major purposes of assessment as found in the literature. The TCoA survey instrument has been used by Brown and colleagues in several backgrounds such as New Zealand (Brown 2002, 2004); Queensland, Australia (Brown 2006, 2008); Hong Kong (Brown, et al. 2011), Spain (Remesal 2011) and Cyprus (Michaelides and Brown 2010) to observe teachers' conceptions towards *negative* conceptions which consist of two main ideas (i.e., assessment is bad, and assessment is ignored) and *positive* conceptions consisting of three main ideas (i.e., assessment improves teaching, assessment improves student learning, and assessment holds schools accountable).

When learning is viewed as continuous development enhanced by structured and meaningful educational experiences, the resulting assessment selection is more likely to yield documentation and feedback associated with the *improvement* belief (Delandhsere and Jones 1999, 219). Improved orientation is thus usually associated with formative assessment in contrast to an *accountability* conception which is usually associated with the term "summative" assessment (Brown 2003). It is worth mentioning here that teachers do not always perceive the purposes related to the summative, and to the formative, as always distinctive (Tillema et al 2011).

The notion of the *student accountability* conception, emphasizes the positive and negative consequences related to students' performance such as graduation, grade retention, grades, and tracking (Brown 2002, 2004). By inspecting inter-correlations among the major conceptions, it appeared that Queensland and New Zealand teachers conceived of student accountability as being irrelevant (i.e., as scores on student accountability increased, so did the values for irrelevance), while school accountability was positively correlated with improvement. In contrast, the Hong Kong teachers had strong positive correlations between improvement and student accountability. There were moderate differences between primary and secondary teachers in New Zealand and Queensland especially around the importance of student accountability. Thus, these led to the conclusion that making students accountable for learning may be considered as an irrelevant purpose of assessment or a positive one based on teachers' background and educational level and the country's particular educational and assessment policies (Harris and Brown 2009).

The third conception of the assessment *accountability of teachers and schools*, underscores society's use of data to determine school and teacher quality (Brown 2004). Delandshere and Jones (1999) conclude that when teachers' assessment view is mainly summative and external in nature, teachers regard assessment as a required means of conveying information to an external audience (Calveric 2010). A paper by Brown and colleagues (2011) involving both primary and secondary government schools in Queensland, resulting in the participation of 1525 teachers, shows that teachers work in a context of low-stakes assessments designed to improve classroom practices or inform central agencies about the quality of the system. On the other hand, New Zealand teachers since 2003 have been given a wide array of voluntary-use, non-centrally controlled, diagnostic, quick feedback, teacher-controlled assessments that have allowed teachers to improve learning in a self-managed manner (Brown et al. 2011). In that context, Brown (2008) reported that New Zealand primary teachers actually associated the use of assessment for improvement with school accountability.

The last notion examined through the TCoA; assessment as *irrelevant,* is representing teachers' view of assessment as unrelated to the work of educators and students (Brown 2003). Typically associated with formal testing, educators who adopt this assessment conception reject assessment due to its perceived harmful impact upon teacher autonomy and professionalism (Brown 2003). Supporters of the irrelevance conception believe assessment detracts from student learning and excludes the inclusion of teachers' intuitive evaluations, student-teacher relationships, and in-depth knowledge of the curriculum and pedagogy (Brown 2003).

The paper by Michaelides and Brown (2010) was the sole attempt in the Cypriot context to investigate teachers' conceptions of assessment. It was a survey undertaken by the former in 2009 with a sample of 249 teachers from all sectors of education. The major findings of this study are that the most popular assessment practices reported were the use of teacher-designed assessments, and written assessments given in class. The least popular practice was the use of standardized tests or large-scale exams. Essay-type questions were not found to be common in Cypriot classrooms. Teachers expressed strong agreement with the use of assessment for improving teaching and learning and endorsed conceptions of assessment for holding students accountable. Assessment for school accountability purposes found only slight support. There was slight disagreement in the belief that assessment is irrelevant to education. Demographic factors did not appear significant in shaping teachers' assessment conceptions. The conceptions teachers hold seem to relate to the practices of assessment they report using (Michaelides 2009). Teachers in Cyprus do not regularly believe that assessment is used for making schools accountable, a reflection of the general educational policy which does not have specific strategies for this purpose (Michaelides and Brown 2010). The TCoA-IIIA measurement model has been found to have an acceptable fit for samples of primary and secondary teachers in both Queensland and New Zealand (Brown 2008). The process for validating the instrument in the Cypriot context was undertaken in 2010 by Dr Michaelides in collaboration with Dr Brown. The New Zealand model was found to be inadmissible for the Cypriot teachers, and after exploratory and confirmatory factor analysis, a hierarchical, intercorrelated model of positive and negative conceptions of assessment was developed.

This section has revealed that, it is difficult to establish main trends and ideas that are universal to teachers; as studies come from several educational backgrounds, levels and countries taking different pathways and thus it is difficult to be comparable or lead to any universal conclusions. An assumption that can be made is that although studies prior to and in the early 2000s show that teachers could not see even regular routine classroom assessment as embedded in the learning procedure, far more positive views incorporating classroom assessment have been developed in recent years. It seems that around 2000 the dominant use of assessment was as a summative tool and teachers tended to separate it from teaching and learning (Black 1993; Black and Wiliam 1998b; Clarke and Gipps 2000; Singh 2000). From research undertaken after the influential study of Black and Wiliam (1998) and subsequent to the emphasis in developing FA (starting) from schools in the UK, by the formal announcements–papers of Ofsted, 1998; DfES, 2003 and QCA, 2003, but also emphasis given to FA in the US, China and Western/European countries, it seems that there is a shift in teachers' views about assessment. As a result, many teachers have developed their practices in far more positive ways (Berry 2004; Brown et al. 2011; Priestley and Sime 2005; Segers and Tillema 2011; Torrance and Pryor 2001).

Also, the literature shows that across countries and sectors, teachers consistently had positive agreement means for both improvement and student accountability conceptions and low means for school accountability and irrelevance conceptions (Harris and Brown 2009). Generally, relationship trends document that those who believe in either the irrelevance or the improvement belief will not traditionally endorse both (Calveric 2010). Also, there are some universal beliefs about assessment purposes that may echo the low-stakes assessment policy. It is suggested that teachers can hold positive and negative conceptions of assessment simultaneously due to the plurality of beliefs and the complexity of the nature of assessment and its many uses. Teachers hold positive and constructive sets of beliefs that empower them to conduct "educational" assessment that is, assessment that leads to improved teaching and learning. They do not by and large believe that assessment is bad or that it should be rejected as being invalid, irrelevant and negative.

The research questions guiding this study are:

1) a. What are Cypriot teachers' conceptions of assessment purposes?
   b. What assessment methods do Cypriot teachers associate with the term assessment?
2) What are teachers' conceptions regarding the development in assessment approaches in the NNC within the Modern Greek language subject?

Beyond the twenty-four questions available from the TCoA IIIA (Appendix A questions 1-24), sixteen more questions were added to the questionnaire to answer the second research question of this study on perceptions about assessment developments within the NNC (Appendix A questions 25-40) whilst six more questions were added to answer the research question (1. b), what do they associate with the term assessment? (Appendix A questions 41-46). A positively packed scale was used as it is considered to give respondents more options in the positive agreement region which allows them to express more finely grained levels of agreement while generating greater discrimination among the various constructs (Brown 2004).

4. Research Methodology

In this study a sequential mixed-method design was used. There are more than forty mixed-method designs reported in the literature (Teddlie and Tashakkori 2009) and typologies of mixed-method designs, while valuable, can never be exhaustive. The design undertaken for this study was one of the major four as reported by Creswell and Plano (2007). The explanatory sequential design which is used here is a two-phase mixed-method design which starts with the collection and analysis of quantitative data followed by the collection and analysis of qualitative data which depend on the results of the first phase for the second to occur. The emphasis is typically on the quantitative data (Creswell and Plano 2007). The quantitative element occurs first and it is used to make decisions about the qualitative phase. In this study a questionnaire instrument (Appendix A) was addressed to teachers in a large sample in order to uncover tendencies in what constitutes assessment according to teachers, the extent to which they agree or disagree with the five competing purposes of assessment as summarized by Brown (TCoA IIIA, 2006, 2009) and conceptions regarding the assessment developments within the NNC.

A qualitative instrument including interviews with the teachers was perceived as essential for the quantitative results to be understood and for more specific information to be gathered in order to enrich the data, address more issues concerning assessment and explain the quantitative results (can be found as Appendix B before their Greek translation). Some of the strengths of this design are that it is quite straightforward because the two methods are conducted and analyzed separately and the final report can be written in two phases making clear which results are occurring from one phase or the other and how they are related in order to have a more holistic view (Creswell and Plano 2007). In this design the same individuals were included in both data collections although the sample size in the qualitative phase was reasonably much smaller.

4.1 Sample

The total number of Greek language teachers in secondary (lower and upper) schools in the last estimation in 2003/2004 was 1578 teachers including full-time and part-time teachers as well as head teachers. The exact number of Greek language teachers in lower secondary public schools is not apparent but it is estimated at around 789 (full-time, part-time and head teachers). It was decided to visit 11 schools from the whole country based on a stratified sample and calculating the time and resources given. Questionnaires were thus given to all Greek language teachers of those 11 schools. There were approximately 13 Greek language teachers in each school. This was supposed to give a total of 11 x 13 = 143 teachers (n=143). This is a proportion of the 18.1% of total Greek language teachers in lower secondary schools in the country. The actual sample, N=95 though, is smaller due to a very low response rate which even led the researcher to visit more schools than planned but again it is quite large for the context.

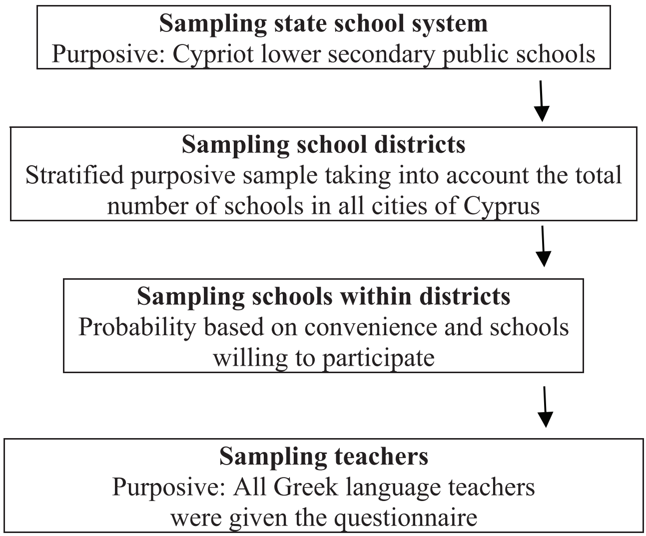


Figure 1: Sampling strategies for quantitative data

Most participants were female (84%), consistent with the overall composition of the Cypriot teacher body where females constitute 81% of the elementary teacher population (Statistical Service 2009).

Table 1: Male and female composition

	Frequency	Percentage
Male	15	15.8
Female	80	84.2
Total	95	100.0

As suggested earlier the number of teachers participating in the qualitative phase of the study was intending to involve people from the same sample selected after the analysis of the quantitative data based on contradictory or extreme conceptions. Only 7 teachers agreed to be interviewed and they were all included in the qualitative phase. This is not an uncommon case according to Teddlie and Tashakkori (2009).

5. Results

The TCoA (IIIA) inventory was used to answer the research question:

1) a. *What are Cypriot teachers' conceptions of assessment purposes?*

To begin with, a summary of the total mean scores for each of the five scales is given (Table 2). The responses were coded based on the original TCoA IIIA as 1 = Strongly disagree, 2 = Mostly disagree, 3 = Slightly agree, 4 = Moderately agree, 5 = Mostly agree, and 6 = Strongly agree.

Table 2: Total mean scores of each scale

	N	Mean	Std. Deviation
irrelevance/ignore	95	2.39	. 982
bad	95	2.43	. 789
improvement of teaching	95	3.45	. 928
improvement of learning	95	3.92	. 942
school accountability	95	3.34	. 907

As might be expected teachers seem to be more agreeable with respect to the positive conceptions as to the three secondary ideas "Assessment improves teaching" (m=3.45), "Assessment improves learning" (m=3.92), and "Assessment holds schools accountable" (m=3.34). Teachers seem to be more inclined to disagree with the negative conceptions "Assessment is irrelevant" (m=2.39) and "Assessment is bad" (m=2.34) showing that they "mostly disagree". These results suggest that in general, teachers value assessment though no strong feelings are apparent. The results of the questionnaire suggest that teachers generally believe assessment to be something good that improves teaching and learning whilst making schools accountable. Teachers show a general disagreement with the notion that assessment is bad or something that is irrelevant which should be ignored.

The results from the questionnaire and the interview data demonstrate complex conceptions regarding assessment as a negative process. The quantitative instrument reveals that the idea of assessment results being filed and ignored is not consistent with teachers' perspectives. Filing and using results are not compulsory in secondary education, and external monitoring policies are uncommon; rather, this rests with the individual teacher and the guidelines communicated by the school principal, who may or may not instruct the teachers to file assessment results in certain formats. The individual teacher may choose, for example, to share student results with parents or present them to an inspector, but that is entirely up to each teacher. The interviews with the teachers reveal that neither everyday assessment nor written tests are regularly used to help teaching and learning.

As far as the notion of making schools accountable is concerned, in the Cypriot context there are no specific policies for making schools accountable through assessment or any other method. The fact that Cypriot Greek language teachers slightly agree with this might be because teachers share a conviction that their assessments are a valid basis for evaluating schools and performance. Teachers seem to argue that because Cyprus is a very small country, the number of schools in each level is not large and teachers teach in several schools during their career, talking to each other about the quality of the schools which is usually translated into assessment results (for example; how many students received an honours degree, how many students failed the final exams and thus failed to proceed to the next level). Also, as rightly stated by the sole previous study undertaken in Cyprus using the TCoA:

> The lack of centralized policies for school accountability and unfamiliarity with unintended consequences of such policies that have been observed in other countries may render this conception attractive given the generally positive stance toward the classroom assessment process. (Michaelides and Brown 2010)

With regard to the idea that assessment should be helpful to teachers differentiating their teaching, teachers argue that it is difficult to implement especially in a mixed ability classroom of approximately 25 students which is considered to be a big number. In addition, they feel that it is up to each teacher whether he/she uses assessment information formatively to help students. A small indication of helping students individually based on their needs is the oral and written feedback but it is not given very often nor are children encouraged to act upon it.

For assessment in improving learning the mean score was M=3.92 showing that teachers generally moderately agree. As the triangulation of data reveals, teachers' answers to the questionnaire show how ideal assessment should be and not what is actually happening in practise. This can sometimes be a completely different story. According to the interviews, teachers believe that the educational system is not very supportive in using alternative methods of assessment or changing the teaching according to students' needs and progress. It is interesting that when trying to support the idea that learning is enhanced through assessment they refer back to the test and that the questions are classified to higher and lower orders for all students to be able to perform well to a certain extent. Some teachers feel that it would have been very constructive for students to do more projects but there is a lack of time both for completing projects and then for giving productive feedback to students.

Following are the results of the RQ:

1) *b. What assessment methods do Cypriot teachers associate with the term assessment?*

With regard to the term "assessment", quantitative data reveal that teachers believe it to be an everyday procedure. They do not feel that assessment is the only test though it is the one given the most importance. What can also be noticed through the triangulation of data is teachers trying to use several methods of assessment; mainly oral and written assessment while they observe students' attitude in the classroom daily, and students undertake projects and assignments and 15 minutes of assessed exercises on the previous lesson. All these approaches though are not regularly used for enhancing teaching and learning but are due to teachers' constant anxiety to be fair as they are solely responsible for showing a child's progress in every level as well as assigning a representative grade and judging if a child is capable of continuing to the next grade. This is consistent with Hasselgreen, Carlsen and Helness (2004) where, in their survey of 914 second language teachers from 37 European countries, they found that, it is more likely for teachers to rely on assessing to award certificates or for placing students rather than using assessment for diagnostic purposes; although these teachers do not see the purpose of assessment as leading to a grade (cited in Vogt and Tsagari 2014).

As for the results of the RQ:

2) *What are teachers' conceptions regarding the development in assessment approaches in the NNC within the Modern Greek Language module?*

One of the actions given emphasis through the NNC is the everyday *oral communication and participation in the classroom* where children can interact with their teachers, and the teacher should make on the spot decisions to support learning. In this area teachers seem disappointed with the contribution of children in building knowledge. Though teachers value oral communication that they argue challenges children and ask them questions while trying to involve them in their learning, students only rarely show the appropriate interest or willingness to participate. They say that students usually ask lower-order questions and teachers usually end up with only a few people in the classroom participating. This may be an important issue to be further developed and discussed and for stakeholders and practitioners to keep in mind. This reform within the NNC might be a great opportunity to train teachers on making their lessons more interesting and for implementing specific strategies for keeping their students' level of interest high and making them contribute to the lesson effectively.

Information from both data instruments but mostly from interviews reveals that teachers do not have a consensus in their minds about *self and peer assessment* methods though these are promoted within the NNC. When they were asked about these activities their answers show ignorance on what they really involve, but nevertheless they have some excuses as to why they do not implement them. They feel that both they as teachers and their students need more training in order to implement these approaches fully. Some peer assessment methods mentioned that are rarely used by teachers are reading to students aloud or giving them written fragments of essays or answers to written tests and projects in order to discuss and find out what they could have done better. Criteria are rarely given explicitly because teachers feel that children are supposed to already know for example what constitutes an excellent essay, or an appropriate answer to a question.

With respect to *feedback*, teachers seem to appreciate both oral and written feedback and acknowledge that it should concern both students' performance and their learning needs. However, the questionnaire data show that feedback should be an everyday process; in practice teachers use it orally only to discuss the written summative assessment when given back to students and not as a daily process. Teachers also give more written feedback on essays rather than tests. The interview data reveal that students rarely take advantage of it in altering the "gap" in learning, thus feedback remains summative in this sense.

With regard to the content of *written assessment* teachers appreciate that it should include both higher and lower-order questions. This was given much attention by the MoEC and the NNC as teachers were given specific examples on how to construct a test and how to involve both higher and lower-order questions. Much emphasis was given by the stakeholders to including unknown tasks in written assessment in order for children to apply what they have learned and be able to present similarities and differences between tasks. This is a specific action mirroring the attempt to promote children in developing critical thinking and using knowledge reflectively.

This study shows that while teachers are trying to implement formative assessment, what is happening in real practice (according to the interviews) is related to what Torrance and Pryor (1998) named "convergent" assessment. "Convergent" assessment can be characterised by three major features 1) assessing IF the learner knows, 2) accomplishing this by closed or pseudo open questions, and 3) at least implicitly this can be characterised as behaviourist, basically attempting to achieve mastery of learning by continuous assessment and not formative assessment. On the contrary, "divergent" assessment can be described as 1) assessing WHAT the learner knows, 2) the use of open questions, and 3) a constructivist view of learning is adopted (Torrance and Pryor 1998).

6. Discussion and Implications

The data reveal that Greek language teachers in Cyprus are invited to play a dual role as per their assessment obligations. One role is as language testers, obliged to construct valid and reliable measurement instruments mostly assigning grades and leading to formal decisions. The other role is being asked to act as teachers who assess their students regularly and use assessment to benefit teaching and learning. This is not uncommon in countries with low-stakes assessment policies where formal, official assessment is not apparent. However, this makes teachers' roles even more challenging. Thus, the researcher suggests that emphasis should be put on two aspects which are not new, but also mentioned and analyzed by several other researchers. These are, a firm terminology on what constitutes assessment and its many purposes and more inclusive and efficient training for pre- and in-service teachers.

Although there have been major efforts in the last twenty years to develop a unified theory of assessment for the educational context that takes into account the culture of learning (Black and Wiliam 2006; Pryor and Crossouard 2008; Shepard 2000), such initiatives still fall somewhat short of the exhaustive and comprehensive theory needed to cover the complex and multidimensional nature of assessment (Simon et al. 2010). Beyond the lack of a firm theory there is limited access to good assessment design and use. What they need is a variety of living and concrete examples of implementation, from which they can derive both conviction and confidence that they can do better and see concrete examples of what doing better means in practice (Black and Wiliam 1998a). This is obvious from teachers' own words in this study's interviews where teachers seem to be willing to modify their assessment practices but gauge concrete examples and guidance. This study also supports the view that more research should be developed about assessment within the Cypriot educational context in order to develop theory as well as practice.

Whilst assessment to be used formatively is a major aspect of the NNC, a lack of firm terminology on FA is still apparent in the curriculum here and elsewhere such as in Priestley and Sime (2005), where it was obvious that teachers perceived formative assessment as enhancing some of the strategies they were already implementing but not actually naming/labelling them. Likewise, in Berry et al. (2019), teachers had shown a lack of confidence in the terminologies (e.g. what is formative assessment?) but when the discussion evolved in the interviews it seems as if they use a wide range of assessment techniques but characterise them as teaching activities. This phenomenon, in Cyprus but also in the United Kingdom, France and Spain (where the Berry et al. 2019 results come from) can be minimized if a common and firm theory is established. Specifically, in this study, it was revealed that though formative assessment is one of the priorities of the NNC the terminology is not widespread in the Cypriot context for unknown reasons. This is why the term was not used in the questionnaires, although when asked in the interviews, teachers seem to slightly recall the terminology "formative assessment" as it was introduced to them in preliminary pre-service training but they are not overly familiar with it. I suggest that it might have been useful for teachers to have a terminology in mind and what it involves. When this terminology was introduced by me in the interviews, teachers said that it involves some of the strategies they were already implementing but without knowing they were called formative assessment; this runs parallel to the research of Priestley and Sime (2005).

The triangulation of data in this study further reveals that teachers are positive about the main elements of FA as identified by Black et al. (2003) as well as the more fundamental role given to assessment related to ongoing instruction similar to studies such as those by Torrance and Pryor (2001), Priestley and Sime (2005) and Berry (2004). For example, the data obtained from the questionnaire show that teachers not only welcome the use of feedback for learning and oral questioning as an assessment process but also their use to develop learning and guide teaching while they also have positive feelings about self and peer assessment. The information obtained from the interviews shows that teachers do not have a consensus in their minds about how to put these strategies into practice in effective ways. This may be attributed to insufficient training and theory not only in the Cypriot context but also apparent elsewhere e.g. Taylor states that regular complaints have been expressed by language testing specialists as to the quality of teacher training programmes in language assessment. There is a perception that even if especially made for this purpose, programmes fail to provide sufficient input on assessment. Taylor (2009, 23) advises that programmes "typically devote little time or attention to assessment theory and practice, perhaps just a short (often optional) module" (cited in Vogt et al. 2018). In the Cypriot context we can see a lack of both systematic research and literature as well as a lack of emphasis given to teachers' professional development in assessment approaches. This second major issue, the lack of sufficient training for in- and pre-service teachers is thus apparent in this study as in many others, such as Taylor (2009) and Lam (2014), whilst the benefits of training are more than apparent in many studies such as in the recent study of Amirian et al. (2016).

Enacting a policy around assessment without the necessary contextual supports is sure to miss the mark. However, a systematic approach that takes key stakeholders–students, teachers, and education leaders into consideration has a stronger chance of success (CEELO 2014; Pinchok and Brandt 2009). In Cyprus, the training that teachers received was not focused on assessment and it was not systematic, nor did it take seriously the role of students as learners in reforming the NNC.

7. References

Amirian, Seyed M. R., Minoo Pourfarhad and Asghar M. Nafchi. 2016. "Teachers' Assessment Literacy and Its Correlation with IELTS Students' Achievement in Writing Skill." *Theory and Practice in Language Studies* 6(5):994-999. DOI: http://dx.doi.org/10.17507/tpls.0605.12.

Askew, Susan. 2000. *Feedback for Learning.* London: Psychology Press.

Berry, Rita. 2004. "Teachers' perception of their roles and their students' roles in the formative assessment process." Paper presented at the Australian Association for Research in Education (AARE), Doing the public Good, Melbourne, Australia, November 2004. Available at: http://www.aare.edu.au/data/publications/2004/ber04978.pdf (Accessed 11 November 2019).

Berry, Vivien, Sheehan Sheehan, and Sonia Munro. 2019. "What does Language Assessment Literacy Mean to Teachers?" *ELT Journal* 73(2):113-123.

Black, Paul, and Dylan Wiliam. 1998. "Assessment and Classroom Learning." *Assessment in Education: Principles, Policy and Practice* 5(1):7-71.

—. 2010. Inside the Black Box: Raising Standards through Classroom Assessment." *Phi Delta Kappan Magazine* 92(1):81-90.

Borg, Simon. 2003. "Teacher Cognition in Language Teaching: A Review of Research on what Language Teachers Think, Know, Believe, and Do." *Language Teaching* 36(2):81-109. Broadfoot, Patricia. 2007. *An introduction to assessment*, New York: Continuum.

Brown, Gavin T. 2004. "Measuring Attitude with Positively Packed Self-report Ratings: Comparison of Agreement and Frequency Scales." *Psychological Reports* 94(1):1015-1024.

Brown, Gavin T. L. 2003. "Teachers' Conceptions of Assessment." PhD Thesis. University of Auckland. Available at: https://researchspace.auckland.ac.nz/handle/2292/63 (Accessed 08 August 2015).

Brown, Gavin T., and Gerrit H. Hirschfeld. 2005. "Secondary School Students' Conceptions of Assessment." *Conceptions of Assessment and Feedback Project Report* 4. Auckland, NZ: University of Auckland.

—. 2008. "Students' Conceptions of Assessment: Links to Outcomes." *Assessment in Education: Principles, Policy & Practice* 15(1):3-17.

Brown, Gavin T., and Gerrit H. Hirschfield. 2007. "Students' Conceptions of Assessment and Mathematics: Self-regulation Raises Achievement." *Australian Journal of Educational and Developmental Psychology* 7:63-74.

Brown, Gavin T., and Lois Harris. 2009. "Unintended Consequences of Using Tests to Improve Learning: How Improvement-oriented Resources Heighten Conceptions of Assessment as School Accountability." *Journal of Multidisciplinary Evaluation* 6(12):68-91.

Brown, Gavin T., Lois R. Harris, and Jennifer A. Harnett. 2012. "Teacher Beliefs about Feedback within an Assessment for Learning Environment: Endorsement of Improved Learning over Student Well-being." *Teaching and Teacher Education* 28(7):968-978.

Brown, Gavin T., Robert Lake, and Gabrielle Matters. 2011. "Queensland Teachers' Conceptions of Assessment: The Impact of Policy Priorities on Teacher Attitudes." *Teaching and Teacher Education* 27(1):210-220.

Brown, Gavin T., S. Earl Irving, Elizabeth R. Peterson, and Gerrit H. Hirschfeld. 2009. "Use of Interactive–informal Assessment Practices: New Zealand Secondary Students' Conceptions of Assessment." *Learning and Instruction* 19(2):97-111.

Brown, Gavin. T. 2006. "Secondary School Students' Conceptions of Assessment: A Survey of Four Schools." Conceptions of Assessment and Feedback Project Report 5. Auckland, NZ: University of Auckland.

—. 2008. *Conceptions of Assessment: Understanding what Assessment Means to Teachers and Students*. New York: Nova Science.

Brown, Gavin. T., Hui, K. F. Sammy, Yu, W. M. Flora, & Kennedy, Kerry. J. 2011 "Teachers' Conceptions of Assessment in Chinese contexts: A tripartite model of accountability, improvement and irrelevance." *International Journal of Educational Research* 50(5-6):307-320.

Butler, Susan M., and Nancy D. McMunn. 2006. *A Teacher's Guide to Classroom Assessment: Understanding and Using Assessment to Improve Student Learning.* San Francisco, CA: Jossey-Bass.

Calveric, Sarah. B. 2010. "Elementary Teachers' Assessment Beliefs and Practices." PhD dissertation. Virginia Commonwealth University. Available at: https://scholarscompass.vcu.edu/cgi/viewcontent.cgi?referer=https://search.yahoo.com/&httpsredir=1&article=3331&context=etd (accessed 1 March 2019).

CEELO. 2014. "Leadership Self-assessment". Accessed March 1st 2019. http://ceelo.org/.

Clark, Christopher M. and Peterson, Penelope L. 198). Teachers' thought processes in Witrock M. C. (Eds.), *Third handbook of research on teaching*, New York: Macmillan, 255-296.

Clarke, Shirley and Gipps, Caroline. 2000. "The role of teachers in teacher assessment in England 1996-1998", *Evaluation & Research in Education* 14(1):38-52.

Creswell, W. John, and Vicki, L. Plano Clark. (2007). *Designing and Conducting Mixed Methods Research*. Thousand Oaks, CA: Sage.

Delandshere, Ginette, and John H. Jones. 1999. "Elementary Teachers' Beliefs about Assessment in Mathematics: A Case of Assessment Paralysis." *Journal of Curriculum and Supervision* 14(3):216-240.

Dixon, Helen, and Mavis Haigh. 2009. "Changing Mathematics Teachers' Conceptions of Assessment and Feedback." *Teacher Development* 13(2):173-186.

Dochy, Filip J. R. C., and Liz McDowell. 1997. "Introduction: Assessment as a Tool for Learning." *Studies in Educational Evaluation* 23(4):279-98.

Duncan, C. Randy, and Brian Noonan. 2007. "Factors Affecting Teachers' Grading and Assessment Practices." *Alberta Journal of Educational Research* 53(1):1-21.

Fulcher, Glenn. 2012. "Assessment Literacy for the Language Classroom." *Language Assessment Quarterly* 9(2):113-132.

Gardner, John. 2012. *Assessment and Learning* (2nd ed.). London: Sage Publications.

Harlen, Wynne and James, Mary 1997. "Assessment and learning: differences and relationships between formative and summative assessment", *Assessment in Education*, 4(3):365-379.

Harris, Lois R., and Gavin T. Brown. 2009. "The Complexity of Teachers' Conceptions of Assessment: Tensions Between the Needs of Schools and students." *Assessment in Education: Principles, Policy & Practice*, 16(3):365-381.

Heritage, Margaret. 2013. *Formative Assessment in Practice: A Process of Inquiry and Action.* Cambridge: Harvard Education Press.

Jeong, Heejeong. 2013. "Defining Assessment Literacy: Is it Different for Language Testers and Non-language Testers?" *Language Testing* 30(3):345-362.

Jones, Alister, and Judy Moreland. 2005. "The Importance of Pedagogical Content Knowledge in Assessment for Learning Practices: A Case-study of a Whole-school Approach." *Curriculum Journal* 16(2):193-206.

Kennedy, K. John, Chan, J. Kin S., Fok Ping K. and Yu, Wai.M. 2008 "Forms of assessment and their potential for enhancing learning: Conceptual and cultural issues," *Educational Research for Policy and Practice*, 7 (3):197-207.

Kyriakides, Leonidas. 1996. "Primary Teachers' Perceptions of Policy for Curriculum Reform in Cyprus with Special Reference to Mathematics." *Mediterranean Journal of Educational Studies* 1(2):77-93.

Lam, Ricky. 2015. "Language assessment training in Hong Kong: Implications for Language Assessment Literacy." *Language Testing* 32(2):169-197.

Malone, Margaret E. 2013. "The Essentials of Assessment Literacy: Contrasts between Testers and Users." *Language Testing* 30(3):329-344.

Michaelides, Michalis. 2009, *A survey of teachers' assessment practices and conceptions on a Cypriot sample.* Poster presented at the 74th Annual Meeting of the Psychometric Society, United Kingdom, 20-24 of July, Cambridge.

National Curriculum. 2010. "Common Questions" (online). Available at: http://www.moec.gov.cy/analytika_programmata/ (Accessed 01 March 2019).

—. 2010. "History of the Configuration of the Curriculum for the Public Schools of Cyprus" (online). Available at: http://www.moec.gov.cy/analytika_programmata/ (Accessed 01 March 2019). Pajares, M. Frank. 1992 "Teachers' Beliefs and Educational Research: Cleaning Up a Messy Construct", Review of Educational Research, 62(3):307-32.

Pat-El, Ron J., Harm Tillema, Mien Segers, and Paul Vedder. 2013. "Validation of Assessment for Learning Questionnaires for Teachers and Students." *British Journal of Educational Psychology* 83(1):98-113.

Pinchok, Nick, and W. Christopher Brandt. 2009. *Connecting Formative Assessment Research to Practice: An Introductory Guide for Educators*. Washington: Learning Point.

Popham, W. James 2009. "Assessment Literacy for Teachers: Faddish or Fundamental?" *Theory into Practice* 48(1):4-11. ERIC: EJ823941.

Popham. W, James W. 2017. *Classroom Assessment: What Teachers Need to Know* (8th Ed.).

Priestley, Mark. and Sime, Daniela. 2005 "Formative assessment for all: a whole-school approach to pedagogic change", The Curriculum Journal, 16(4):475-492.

Remesal, Ana. 2011. "Primary and Secondary Teachers' Conceptions of Assessment: A Qualitative Study." *Teaching and Teacher Education* 27(2):472-482.

Sadler, D. Royce. 1989. "Formative Assessment and the Design of Instructional Systems." *Instructional Science* 18(2): 119-44.

Sato, Mistilina, J. E. Coffey, and S. Moorthy. 2005. "Two Teachers Making Assessment for Learning their Own." *The Curriculum Journal* 16:77-192.

Segers, Mien, and Harm, Tillema. (2011). How do Dutch secondary teachers and students conceive the purpose of assessment? Studies in Educational Evaluation, 37, 49-54.

Shepard, Lorrie A. 2000. "The Role of Assessment in a Learning Culture." *Educational Researcher* 29(7):4-14.

Singh, Birendra 2000. "Formative assessment: which way now?", Paper presented at the British Educational Research Association Annual Conference (BERA), 2-5 of September, University of Sussex at Brighton.

Teddlie, Charles and Abbas, Tashakkori. (2009). *Foundations of Mixed Methods Research: Integrating Quantitative and Qualitative Approaches in the Social and Behavioral Sciences*. Sage, London.

Tillema, Harm, Martinj Leenknecht, and Mien Segers. 2011. "Assessing Assessment Quality: Criteria for Quality Assurance in Design of (Peer) Assessment for Learning – A Review of Research Studies." Studies in Educational Evaluation 37(1):25–34. https://doi.org/10.1016/j.stueduc.2011.03.004

Torrance, Harry, and John Pryor. 1998. *Investigating Formative Assessment: Teaching, Learning and Assessment in the Classroom*. Taylor & Francis, Inc.

Vogt, Karin, and Dina Tsagari. 2014. "Assessment Literacy of Foreign Language Teachers: Findings of a European Study." *Language Assessment Quarterly* 11(4):374-402. DOI: 10.1080/15434303.2014.960046.

Vogt, Karin, Ines, Sperling, and Moritz Brüstle. 2018. "TALE–Teachers' Assessment Literacy Enhancement Needs Analysis Report." Erasmus+ European Union. Project No. 2015-1CY01-KA201-011863 (Project No. 2015-1-CY01-KA201-011863). Retrieved from: http://taleproject.eu/mod/page/view.php?id=1198.

Wiliam, Dylan. 2011. "What is Assessment for Learning?" *Studies in Educational Evaluation* 37(1):3-14.

Appendices

Appendix A

Teachers' Questionnaire before the Greek Translation

PLEASE TICK THE ANSWER THAT REPRESENTS YOU THE BEST	STRONGLY DISAGREE	DISAGREE SLIGHTLY	SLIGHTLY AGREE	MODERATELY AGREE	MOSTLY AGREE	STRONGLY AGREE
1) Assessment is an imprecise process	O	O	O	O	O	O
2) Assessment is unfair to students	O	O	O	O	O	O
3) Assessment interferes with teaching	O	O	O	O	O	O
4) Assessment results are filed and ignored	O	O	O	O	O	O
5) Assessment results are consistent	O	O	O	O	O	O
6) Assessment provides information on how well schools are doing	O	O	O	O	O	O
7) Assessment results are trustworthy	O	O	O	O	O	O
8) Assessment helps students improve their learning	O	O	O	O	O	O
9) Assessment allows different students to get different instruction	O	O	O	O	O	O
10) Assessment information modifies ongoing teaching of students	O	O	O	O	O	O
11) Assessment is integrated with teaching practice	O	O	O	O	O	O
12) Assessment results can be depended on	O	O	O	O	O	O
13) Teachers conduct assessment but make little use of the results	O	O	O	O	O	O
14) Assessment measures students' higher-order thinking skills	O	O	O	O	O	O
15) Assessment provides feedback to students about their performance	O	O	O	O	O	O
16) Assessment determines if students meet qualification standards	O	O	O	O	O	O
17) Assessment is a good way to	O	O	O	O	O	O

evaluate schools						
18) Assessment has little impact on teaching	O	O	O	O	O	O
19) Assessment is assigning a grade or level to student work	O	O	O	O	O	O
20) Assessment is an accurate indicator of schools' quality	O	O	O	O	O	O
21) Assessment forces teachers to teach in a way against their beliefs	O	O	O	O	O	O
22) Assessment establishes what students have learned	O	O	O	O	O	O
23) Assessment feeds back to students their learning needs	O	O	O	O	O	O
24) Assessment is a way to determine how much students have learned from teaching	O	O	O	O	O	O
25) Written assessment should involve only higher-order questions	O	O	O	O	O	O
26) Written assessment should involve only lower-order questions	O	O	O	O	O	O
27) Written assessment should involve both higher- and lower-order questions	O	O	O	O	O	O
28) Students should be encouraged to ask questions	O	O	O	O	O	O
29) Teachers should give time to students in the classroom to answer a question	O	O	O	O	O	O
30) Oral questions are a way of everyday assessment	O	O	O	O	O	O
31) Oral questions should involve only higher-order questions	O	O	O	O	O	O
32) Oral questions should involve only lower-order questions	O	O	O	O	O	O
33) Oral questions should involve both higher- and lower-order questions	O	O	O	O	O	O
34) Students should assess each other's work (peer assessment)	O	O	O	O	O	O
35) Feedback should only be about ego involving evaluation	O	O	O	O	O	O
36) Feedback should only be	O	O	O	O	O	O

concerned with the task						
37) Feedback should only be given in written assessment	O	O	O	O	O	O
38) Feedback should be an everyday procedure	O	O	O	O	O	O
39) The procedure where students find the elements that constitute a certain kind of subject document, write it and assess it should be applied	O	O	O	O	O	O
40) The procedure where students reflect on their past work by evaluating it and identify what still needs to be pursued should be applied	O	O	O	O	O	O
41) Assessment means a standardized test	O	O	O	O	O	O
42) Assessment means a teacher made written test	O	O	O	O	O	O
43) Assessment is a formal process	O	O	O	O	O	O
44) Assessment happens every day with oral questions and answers	O	O	O	O	O	O
45) Assessment is happening every day by observation	O	O	O	O	O	O
46) Assessment is an informal procedure	O	O	O	O	O	O

Please give some personal information which will remain strictly confidential:

School Name:

....

I have taught for ……. years

Gender: (circle) male/female

How long have you been trained for the NNC, please answer in days

....

Have you been using the NNC in the school year 2010-2011? (circle) yes/no

Did you participate in any other way in the NNC? Please explain:

....

....

....

If you are chosen for a twenty-minute interview are you willing to participate? (circle) yes/no

If yes, please give your contact details:

....

Appendix B

Teachers' Interviews (outline, before the Greek translation)

According to the findings of the questionnaires and each teacher's answers questions were changed accordingly.

1. What does the term assessment mean to you?
2. The main skills and competencies that children need to develop for the 21st century according to the Cypriot stakeholders are for children to be creative, have critical thinking, use the knowledge reflectively and have interpersonal skills as well as to co-operate with others. How are each of these skills developed?
   Let's talk first about creativity: Critical thinking; Use Knowledge reflectively; Have interpersonal skills and cooperate with others;
   And what is the role of assessment for developing these skills (if any)? (this question might be asked consequently after the teacher elaborates on each of the skills above).
3. What assessment methods do you usually undertake in the Greek language lesson? (this is an indirect question trying to elicit information on formal-government obliged assessment and any informal-teacher assessment. If the teachers do not seem to elaborate on this, they are going to be asked more specifically about the formal assessment they are asked by the government to undertake once every

semester and if any other forms of informal assessment are taking place).

4. For what purposes are each of the assessment methods referred to above used?
5. (If the teacher's answers on the questionnaire show a belief that assessment improves teaching and learning they are going to be asked to elaborate on this and explain how.)
   According to your answers to the questionnaire, you seem to believe that assessment improves teaching and learning. So, can you elaborate on this?
6. (If the teacher's answers on the questionnaire show a belief that assessment should be rejected either because it is invalid, irrelevant or negative.) If they have shown positivity in this belief they are going to be asked to elaborate on this.
7. Do you believe that assessment is separated from teaching and instruction or is it a part of it? Please explain.
8. One of the reforms of the NNC was to divide the semesters into two instead of three. To what extent do you think the division of semesters into two has helped you in using alternative methods of assessment?
9. Considering the training for the NNC that you have undertaken, can you tell me what assessment methods were suggested by the trainer for the Greek Language specifically and if you are for or against them and why?
10. Did the training help you develop your assessment skills, and if yes, how?
11. a) What do you think is the purpose of oral questioning? b) What is the purpose of written questions? c) What do you think about students asking questions? (Do you encourage them?)
12. What do you think about peer- and self-assessment methods? To what extent do you use them and how? Is the NNC supportive of these elements of assessment? Have you been trained for such actions? What do you think is the purpose of feedback and how often do you use it? Is the NNC supportive of this element of assessment? Have you been trained for how and when to use it?
13. Are you familiar with the term formative assessment and what it includes?

CHAPTER FIVE

HOW MUCH DO WE REALLY KNOW ABOUT ALTERNATIVE ASSESSMENT? INVESTIGATING YL TEACHERS' PERCEPTION OF ASSESSING THROUGH GAMES

CHRISTINA NICOLE GIANNIKAS

Age-appropriate and effective assessment approaches have been a matter of debate in the field of language learning and teaching. Nonetheless, when assessing YLs, practitioners tend to fall back on approaches favoured in classes of older learners. The present chapter focuses on the use of games as a means of alternative assessment and draws conclusions from data gathered via a global online teacher's awareness survey. The chapter aims to display the findings of Language Assessment Literacy (LAL) regarding the use of games as an assessment tool, teachers' perspectives, and participants' YL assessment training. The outcomes show the need for the enhancement of LAL, and the need to include alternative assessment in teacher training programmes. Issues emerging from the research, such as doubts as to effective assessment via games, classroom management, and parental involvement will be discussed. The chapter aims to raise awareness of assessment approaches in the YL classroom, of teachers' LAL and perspectives, and their YL assessment training. A further intention of the chapter is to use the findings of the study to provide the empirically derived content of LAL and inform the design and structure of new materials that can be applied in language assessment programmes, focused on YLs' needs.

1. Introduction

Young Learners (YLs) need to be taught differently from their older counterparts; nowadays, teachers are equipped with various creative and entertaining materials and gadgets to assist in making their lessons age-appropriate and motivating. Nonetheless, when assessing YLs, practitioners tend to fall back on approaches favoured in classes of older learners. The general assessment literature has characterized this as the "stable state" of the adult mind (Bialystock 2001). However, McKay (2006) argues that children are in a state of continuous cognitive, social, emotional and physical growth, with very dissimilar traits to adult learners. Ioannou-Georgiou and Pavlou (2003) suggest that adopting similar assessment practices to those seen in adult language classrooms can damage the YL teacher's efforts in encouraging students' positive attitude towards the foreign language. Resorting to "traditional tests" may not always be the ideal solution for children, as they can have negative effects on their self-esteem, motivation, and overall attitude towards the target language. For YL assessment to be effective, it needs to be tied to the principles of learning (McKay 2009), *ergo* the integration of alternative assessment approaches. Such approaches underlie pedagogic principles and can be beneficial and meaningful to the specific age group. In recent years, several forms of alternative assessment have gained a considerable place in the language teaching process (Torres Soriano 2009). These controversial ways of assessment require YLs to demonstrate what they have learned and how they can apply it over a period of time. They can closely mirror teaching approaches which include puzzles, songs, drama, drawing and games.

The present chapter will focus on the use of alternative assessment; more specifically, the chapter draws conclusions from data gathered via a teacher's awareness survey on the use of games as an assessment tool in the YL classroom. The chapter aims to present data on the teachers' perspective of assessing young learners through games on a global scale. A further intention of the present chapter is to use the findings of the study to provide empirically derived content of language assessment literacy (LAL) and inform the design and structure of new materials that can be applied in language assessment programmes, focused on YLs' needs.

2. Theoretical Framework

The issue of YLs' assessment has been a matter of great interest in the field of applied linguistics, and the need to apply forms of age-appropriate assessment has grown. Research has shown that traditional classroom testing procedures can cause YLs anxiety that affects their progress, self-image and L2 development (Smith 1996). Conner (2008) argues that poorly formed and administered paper-and -pencil tests can have a negative effect on learning and lead to losing interest in language learning. Therefore, it is vital that children are evaluated in a stress-reduced environment, to help them perceive the assessment procedure as an integral component of the learning and teaching process (Bejarano and Gordon 2009). Hasselgreen (2005) has argued that when assessing children, practitioners need to consider their students' maturity and present them with tasks that provide frequent language input. This has led to many educational reforms, which have declared it is time for assessment approaches that go beyond the traditional paper-and-pencil techniques (Stiggins 1997) and on to alternative assessment approaches, intended to encourage the use of authentic activities, enrich the language learning experience and motivate students to take responsibility for their language learning (Wolf, Bixby, Glenn and Gardner 1991).

Alternative assessment techniques present a dynamic of YLs' linguistic development, rather than a static picture (Shaaban 2005), and call for needing more than one way to measure students' progress (Butler 1997). Assessment should be an ongoing process which measures the students' ability to apply knowledge and skills in authentic contexts. Nonetheless, the attempted change in education does not necessarily mean that the assessment concept within teachers' practice has also changed. Although alternative assessment mirrors more familiar and realistic approaches to assessing children (Torrance 1998; Ontario Ministry of Education and Training 1995), there have been advocates for its implementation. Shute (2006) has argued that learning in schools is still focused on the acquisition of content within a teacher-centred model. Certain implications have led to the conclusion that “teach-to-test” approaches do not deliver the desired outcomes in learning processes as students' short-term motivation disappears well before the completion of their language learning course (Su 2000). The difference that alternative assessment makes is that it requires teaching skills, which allow practitioners to evaluate students' knowledge, record their progress, provide effective feedback in the teaching process and mend their practice so as to meet students' needs, depending on what the assessment outcomes reveal (Wiggins 1989).

Inbar-Lourie and Shohamy's (2009) research on YLs and alternative assessment has led them to propose an assessment construct that is grounded in a view of the L2 as a tool for gaining knowledge, meaning and cognitive processing skills. The innovative methods Inbar-Lourie and Shohamy propose are considered to focus on the spirit of the child, and activities such as games are intrinsically motivating and intriguing to this particular age group (Ioannou-Georgiou and Pavlou 2003). Klopfer, Osterweil and Salen (2009) distinguish games as structured or organized play, which enables cognitive and physical engagement that allows for the freedom to fail, recover and experiment, and apply effort and interpretation. In a similar tone, Prensky (2001) has stated that games consist of a number of key elements, such as rules, goals, outcomes, feedback, conflict, interaction and representation. Young students can use their spontaneity in L2 use via games and engage in "real tasks" as the teacher monitors the process and provides YLs with useful feedback (Torrance 1998). Research has shown that games increase children's attention span and improve the efficiency of thinking and problem-solving (Goldstein 2012). When playing games, YLs have the opportunity to put the language they are learning into a meaningful context and develop the ability to understand new concepts or ideas, consider different perspectives and, most importantly, experiment with the language (Giannikas 2017). Vernon (2006) suggests that when children play a game they concentrate on the task and absorb the L2 subconsciously, while Yolagelili and Arikan's (2011) research shows that games are an important and necessary part of English Language Teaching (ELT), simply because they provide practitioners with many instructional advantages. Given the nature of the specific form of assessment, and in order to execute the teaching techniques required, teachers need a high level of LAL. Although alternative assessment has been widely accepted as a tool of educational reform, a lack of LAL among teachers and school leaders is a commonly encountered problem when attempting to apply alternative assessment (Koh 2011).

Based on the outcomes of Tsagari and Vogt's (2017) Europe-based study, teacher education programmes do not provide adequate training in LTA; a shortcoming of professional development in language teaching which drives teachers to use non-risk published assessment material. Participants of the same study indicated that they were aware of alternative forms of assessment, but did not include them in their assessment routine. In particular, findings showed that primary school English teachers in Greece were not warm to the idea of implementing alternative forms of assessment. A few dedicated teachers in the German context were more willing to use innovative forms of assessment but would terminate any form of alternative assessment if they ever had a negative experience applying it in the classroom. Studies such as Tsagari and Vogt's show the lack of knowledge of designing, developing and evaluating assessment procedures in the traditional sense (Vogt and Tsagari 2014). According to Stiggins (2002), in most North American jurisdictions, there is relatively little emphasis on assessment in teacher education. More specifically, out of 10 Canadian provinces and 50 US states, only Hawaii and Nebraska currently invest a significant amount of capital targeted at assessment and evaluation practices. State evaluations have suggested that professional development tied to Nebraska's School-based Teacher-led Assessment Reporting System (STARS) has had a positive impact on teacher confidence, knowledge, and skill in classroom assessment (Bandalos 2004;

Lukin 2004). Nonetheless, there is still an extended need for LAL at a global scale.

Green (2016) argues that the development of alternative theoretical approaches makes it possible to offer LAL courses for language teachers that escape the language testing tradition. The teacher's role to apply principles of sound assessment (Popham 2004) is vital. This role has come under scrutiny in the field as a result of using assessment for accountability purposes (Leung 2014a), as well as policies to integrate assessment-for-learning principles in the language learning curriculum and L2 assessment (Fulcher 2012). To apply alternative assessment in any shape or form, language teachers need a solid background in this area in order to be well-equipped, and able to implement assessment techniques to also utilize appropriate forms of teaching and meet students' needs (McMillan 2000). As teachers are key agents in educational assessment (Leung 2014b), LAL has been of growing interest in the field for quite some time; partly due to its importance in student learning (Black and William 1998), which means that embracing learning via assessment could be most effective in the language education context of the YL (Giannikas 2017).

3. Research Methodology

The needs of the current study demanded a global reach in order to collect data adequate to the aims of the investigation. This was accomplished through an online survey, which was selected due to the method of contact, the medium and the responding mode (Tourangeau, et al. 2000; Tsagari and Giannikas, in preparation). Online surveys allow for the collection of data on an array of issues regarding the behaviour, thoughts, and feelings of the target participants. The purposes of survey research include describing a specific population, identifying the population's characteristics, describing attributes of research interest, explaining a phenomenon or how variables are related. Survey data can be either quantitative in nature with statistical outcomes, or qualitative with detailed narrative outcomes (Buchanan and Hvizdak 2009).

The present survey was created on Google Forms, and delivered via emails and private social media platforms. The intended population was described as "language teachers" and respondents were estimated to spend 20-25 minutes on the survey. Responses were automatically loaded into a database on the web server, from which they could be downloaded on Excel. The data were automatically stored using the Google Forms database technology and an HTML.

The survey consisted of a series of open- and closed-ended questions in five broad areas:

1) Background information,
2) Teacher education,
3) The use of play in the language classroom,
4) Assessing YLs through games, and
5) Respondents' perspective of LAL.

In the analysis of both the quantitative and qualitative responses, data reduction and coding strategies were applied. The closed responses were analyzed by using an exploratory factor analysis, whereas the qualitative responses were summarized and coded. The coding matrix emerged through the manual collocation of categories that emerged (Tsagari and Giannikas, in preparation). The analysis of the survey data attempted to identify the use of play and the components of LAL.

4. Findings

4.1 The Participants

The online survey collected 74 responses from language teachers based in Cyprus, Greece, Portugal, France, Sweden, Brazil, Mexico, Canada, Italy, and Spain.

According to the responses, a high percentage of the participants worked in the private sector (85.5%). Data showed that 27.5% of the respondents had been teaching YLs for 1-5 years, 19.4% had been teaching YLs for 6-10 years, and only 9.7% of the respondents had been teaching English to young children for over 30 years. The data also showed that the majority of the practitioners preferred teaching older YL groups (87.5% teach 8-10 year-olds and 84.4% teach 11-12 year-olds vs. 57.8% teach 5-7 year-olds and only 20.3% teach 3-4 year-olds).

4.2 Participants' Pre-/In-Service Training

The training areas in the questionnaires focused on general training, on teaching young learners, as well as on the integration of alternative assessment. This was to display the foundation that teachers have in their general teaching, in addition to their LAL. The data regarding general training from each participating country are displayed as percentages in Table 1:

Table 1: General Training on Teaching YLs

Countries/Training	Pre-service Training	In-service Training
Greece	52%	60%
Cyprus	75%	50%
Italy	0%	0%
Spain	80%	100%
France	0%	100%
Portugal	20%	100%
Sweden	100%	0%
Canada	100%	100%
Brazil	80%	0%
Mexico	80%	100%

According to Table 1, there is no clear policy on pre- or in-service training in Europe. In countries such as Greece, Cyprus and Spain teachers are expected to undergo pre- and in-service training almost equally. Nonetheless, the data show that higher percentages of teachers undergo in-service training in Greece and Spain, whereas more teachers are expected to undergo pre-service rather than in-service training in Cyprus. There are countries where teachers are not supported with any in-service training, such as France, and others that were not supported with any training at all, such as Italy. In Portugal, there is a disproportionate balance of 100% of respondents who stated that they had undergone in-service training but only 20% of the teachers had pre-service training, as well as in Sweden where all teachers had pre-service training but no in-service training. Respondents from Canada stated that 100% had pre- and in-service training. Mexico comes close with similar percentages of 80% of the respondents having undergone pre-service training and 100% had done in-service training. Teachers based in Brazil stated that 80% of them had pre-service training, however none of them had done any in-service training. The participating teachers were asked to mention the reasons they had no pre-service or in-service training, and the most popular responses were: 1) none was available, 2) lack of funding, 3) lack of time, 4) lack of interest, and 5) private school owner who was self-trained.

4.3 The Use of Play in Teaching

The next area that the study concentrated on was the use of games for language learning which prompted data that showed the teachers' experience and understanding of using games in their practice. This set of standards establishes a benchmark of expectations for LAL and assessing YLs through games. The data display the foundational underpinnings, which need to be examined before we can effectively investigate the issue of LAL and alternative assessment. The teachers who participated in the study were asked to scale the importance of games in the language classroom and the use of play in their teaching. The majority of the teachers stated that games are very important in L2 development. The majority (65.6%) also stated that they implement games in their language lessons, and only 3.1% do not. Even though most of the respondents use games in their teaching, the percentage is low for teachers of YLs. There were teachers who stated that they sometimes used games (33.3%) but could not integrate games into their daily practice due to either lack of time (75%), classroom management issues (50%), or in order to save time and cover a coursebook by the end of the school year (75%). The teachers were also requested to specify the benefits playing games had on their YLs' language learning. Certain themes emerged from the coding of the data, and the most common responses are displayed in Table 2:

Table 2: The benefits of play in language learning

Themes	Percentages
Creativity	12%
Engagement	8%
Motivation	14%
Entertainment	30%
Acquiring the L2 with ease	8%
Less stressful	10%
Focus	10%
Cooperation and Interaction	10%

Of all the benefits mentioned in the online survey, the issue of entertainment is, clearly, considered the most important and effective reason to use games in language teaching with children, according to the informants (30%). The second theme that emerged, based on the teachers' responses (14%) was an increase of motivation observed when games were implemented in the lesson/curriculum; and a close third was the enhancement of creativity (12%). The following extracts offer a deeper understanding of the respondents' perspective:

> **Teacher 15:** Young learners react better to learning in a fun environment, they create happy memories and leave the classroom energised and excited, eager to come back for more. It is also a more student-friendly environment where they can interact with other children using the target language.

> **Teacher 21:** Of course, there are a lot of benefits. Real life is how babies and children learn their native language, so using play helps them learn the new language in a more spontaneous, fun and practical way. As a result, they enjoy it much more and the effect on their command is long lasting.

> **Teacher 22:** Games benefit children in the sense that they allow them to use English in a natural and creative way.

> **Teacher 64**: Children learn more easily and naturally, they like playing roles, they like trying to win, so they become more eager and motivated. This is how they observe knowledge at these ages.

The data analysis also uncovered themes that prompted limitations in the use of play for language learning. The themes that emerged indicated that lack of training brought limitations to applying one of the most evident teaching tools in the YL classroom. The main limitations were: 1) that games are ideal for kinaesthetic learning, which limits learning only to games that involve movement, and 2) outdoor learning, which limits playing games to teachers who have space outdoors and during certain seasons of the school year. These limitations raise questions as to the role that games play in language teaching, according to practitioners, and whether a tool that encourages high levels of linguistic and cognitive development is overseen or underused in the YL classroom.

4.4 Assessing YLs via Games

The fourth part of the online survey relates to implementing assessment through games and the opportunities, as well as the challenges, that teachers have experienced or foresee when applying games in their assessment routine. According to the data, 45.3% of the informants use games as an assessment tool, which is 20% less than when teachers were asked if they implement games in their teaching. A higher number of 23.4% stated that they do not use games for assessment purposes, whereas 31.3% stated that they sometimes use games to assess their YLs.

There were more negative responses as to why games would not be beneficial in the assessment process than in the teaching process. Half of the teachers were reluctant to assess their YLs via games because they were not familiar with techniques they could apply in order to keep a record of the students' performance and progress. A large number of 45% of informants also raised the issue of classroom management, a similar number to why teachers do not use games in their teaching either (50%). Another concern of 35% of the teachers was that children would not concentrate on the task at hand and would view games as "fun time" and not as part of their assessment. Additionally, 20% of the respondents stated that assessment is an issue that is taken very seriously at their school. Assessing children through games rather than the traditional pencil-and-paper method would create tension between the teacher and the principal. Similarly, 20% of the respondents were concerned that the integration of games in the assessment process would prompt parents to complain about the said teacher's assessment method. Finally, 25% of the informants were not convinced that games as an assessment tool would work or have any positive effect on their learners. Nonetheless, there were teachers who were positive towards the use of games as an assessment tool. When asked what they assess through games the vast majority (97.6%) of the teachers claimed to assess vocabulary, 73.8% use games to assess grammar, 71.4% to assess language comprehension and only 26.2% assessed syntax. Out of the four skills, 69% of the teachers use games to assess listening, 38.1% to assess reading, and 31% to assess writing. Speaking skills were not mentioned by any of the respondents. According to these data, the use of games for assessment purposes has a traditional take to it, as the focus was on vocabulary, grammar and language comprehension, which reflect the focus of the early days of the Grammar Translation Method (GTM). Among the four skills, games were mostly used to assess listening, which is the most straightforward and simple use of a game. Nonetheless, it is challenging to assess and keep an accurate record if LAL is not at a high level. Teachers need to be able to distinguish listening from imitating their peers and able to use their own judgement when giving the students feedback or scores. Therefore, a vital part of assessing YLs through games is to keep valid records during or immediately after the process. When informants were asked about their record-keeping, they gave a number of responses. The themes that emerged from the teachers' responses can be found in Table 3:

Table 3: Techniques or record-keeping when assessing via games

Themes	Percentages
Note-taking after the lesson	80%
Digitally	45%
I do not keep records	30%
Portfolios	15%
Checklist	10%
Report Cards	15%
From Memory	10%
From the projects they produce	10%
It depends on the game	5%

According to the data, the majority of the teachers (80%) take notes after the lesson is complete, which means that games could be played at any time during the lesson. The time period between the game and note-taking was not specified in the survey. However, 42% of the teachers stated that notes were not taken consistently, while 25% of the respondents made notes in order to add to the students' overall grades, however, the feedback was not communicated to the students. Similarly, 15% of the teachers claimed to add the students' performance to their report cards without keeping other records of their assessed game playing *per se*. Digital record-keeping was preferred by 45% of the respondents, saved on a PC after the completion of the lesson. Nearly 30% of the teachers stated that they do not keep records when assessing YLs via games. Half of these teachers indicated that they did not wish to give additional stress to the students by keeping records, and did not deem this form of assessment as "formal" in order to make a record of the students' performance and progress. This immediately eliminates the point of assessing YLs via games as no goals are met, the benefits of assessing YLs are not realized and in reality, there is no assessment involved. A low 10% stated that they keep a record of students' performance and progress from memory, which can be interpreted as no record-keeping at all. Checklists and portfolios were mentioned by a smaller percentage (10% respectively), and these are positive indicators that the specific teachers had knowledge and/or experience of applying games for assessment purposes. Finally, a smaller (5%), yet positive response, was that teachers did not use one specific record-keeping approach, as it depends on the game played and the assessment purposes.

4.5 Respondents' perspective of LAL

The fifth and final section of the online survey focused on the informants' perspective of LAL development. This final part of the survey was pivotal to the study, as it gives insights into the teachers' understanding of what they need in order to develop alternative assessment in LAL with a focus on games and YLs. Table 4 displays samples of the teachers' responses.

Table 4: Teachers' perspective on LAL and training programmes

We need innovative ideas for assessment.
I would like to learn more about informal assessment combined with child psychology.
Practical assessment workshops.
We need practical advice.
Real hands-on seminars and programmes, not only theories, true ideas on how to work with specific aspects to assess young learners.
I think teacher training programmes need to focus on ways to approach children, on how to deal with specific learning abilities and how to create a fun creative environment. This needs to be done before we can even consider thinking about assessment.
I have no teacher training experience so it is impossible to say.

I can't really answer this, as I believe that teacher training programmes are not the only options a teacher has. Here in Greece, many such programmes include using play, but sharing with colleagues is also a great option, as well as doing some personal research. In addition, many language books provide great material too.
I do not assess my students by using games, they are assessed when they take tests. This is the most effective and training is not needed for this.
We need to be shown how to use newer assessment material.

As seen in the sample responses, the needs and awareness of the issue vary. While some teachers yearn for innovative ideas, hands-on training and new material, others find it difficult to rely on training as they feel that peer interaction and coursebooks provide an adequate amount of information. Additionally, there were responses that supported test taking with YLs and no training is needed for this approach. Teachers have identified topics where they would need to know more about assessment and seek training opportunities. Many of the teachers have asked for a practical application of how to implement alternative assessment, and 62.5% have stated that they would attend teacher training on assessing YLs via games; 28.1% were uncertain if they would devote their time to such training, and 9.4% were negative about it. These respondents believe that training on the specific element of language learning would be unnecessary, as the most effective approach to assessing YLs would be test taking.

5. Discussion

In order to collect data that would provide a holistic view of the implementation of alternative assessment in YL programmes, teachers' general training was brought to the surface. The data show that the informants' teacher training, or lack of it, creates obstacles in their teaching as well as their assessment skills. The teacher training policies in the countries mentioned in this chapter vary, with some supporting pre-service training but not in-service, and vice versa, others supporting both, while some do not support either. This implies that there are language teachers across the globe that may enter the classroom without being sufficiently prepared to teach YLs, or they have not been prompted to improve their efficiency, ability or knowledge in their practice. Such gaps in the foundation of teacher education and professional development make it unreasonable to expect teachers to have developed LAL, or be in a position to implement alternative assessment effectively and use play as an assessment tool. Although the majority of the teachers have claimed to use play in their teaching, there were responses which indicated limitations in creating an environment where rich play for effective language teaching and learning can occur. This could lead to the conclusion that the term "playing games in the language classroom" is vague, as it becomes unclear when put in the classroom context, as there seem to be many interpretations of the term.

The play situation in language teaching has led to even more limitations and complications when it comes to using games as an assessment tool. The data of the present study show that teachers are reluctant to apply games in their assessment routine as this may create classroom management issues, prevent YLs from concentrating, and conflicts may occur with parents and headteachers. This raises the same question that Taylor (2009) had raised concerning balance among the technical know-how, the practical skills, the theoretical knowledge and the context of assessment within the educational system and society. The present findings reveal that YL teachers have received inadequate training in the fundamentals of assessment practices, and endorse the need for teacher training to equip them with the skills to apply alternative assessment. It is important for teacher candidates, as well as in-service teachers, to develop a mindset that acknowledges the need to evaluate their use of games in assessment. A high percentage of the teachers in this study have stated that they would attend training courses on alternative assessment with a focus on games, which would give them the chance to enhance their knowledge of assessment and the preparation of alternative assessment tasks. Teachers who have opposed such training have done so due to their disbelief in the assessment method and its validity, reliability and objectivity. With the necessary provision, YL teachers can acquire the skills to apply games in their assessment routine and to record YLs' progress and performance during the process. Such programmes would provide a higher level of awareness and self-awareness through LAL for teachers to assess for meaningful language learning (Scarino 2017). It is vital that teachers across the globe become familiar with good assessment practices, which will assist them in determining the appropriateness of the games (or any content for that matter), the pace, and how to monitor students' L2 production and assess their progress. Furthermore, an important factor of good assessment practices is to develop an interpretive nature when assessing via games and when integrating the approach in the classroom. Only if teachers' LAL is developed and practitioners have a clear understanding of the approach will alternative assessment be in any way effective.

6. Conclusion

The chapter aimed to raise awareness of assessment approaches in the YL classroom, of teachers' LAL and perspectives and their YL assessment training. A further intention of the present chapter was to use the findings of the study to provide the empirically derived content of LAL and inform the design and structure of new materials that can be applied in language assessment programmes, focused on YLs' needs. The findings of the study show that the majority of YL teachers across the globe are eager to apply different interactive methods, in order to embrace the characteristics of a young child. As in teaching, not all assessment approaches can be applied and delivered across all age groups. While this chapter is not suggesting eliminating formative or summative assessment with younger age groups, it is urging language practitioners, teacher trainers and designers of teacher training programmes to embrace alternative assessment, and the type of assessment that a young child can relate to. Continuous testing can provoke anxiety and self-doubt in YLs, which is why a balance of assessment approaches and principles is needed for effective assessment. The extent of this would not only benefit children, but also bring long-term professional well-being to teachers (Popham 2009).

While there are studies that have focused on alternative assessment, assessing YLs via games has been overlooked in the literature. There is a need to view assessment as not only a means of grading, but as an element of meaningful teaching. Age-appropriate approaches can help language teachers accomplish efficient alternative assessment. This needs to be investigated further, in order for a substantial contribution to the field of LAL to be made. Although the present study has collected a considerable amount of data, it is not free of limitations. The study would have more in-depth information if it had more sources. Furthermore, the research focused only on practitioners and their perceptions of applying assessment via games and their LAL. Further research could benefit from the input of teacher trainers and training programme designers, who would be able to provide insights from a different perspective. There is a strong need for further research in order to enrich the literature with information that would give merit to YL assessment and help raise awareness of the need for higher levels of LAL among language-teaching professionals via teacher training. Further research can develop a stronger understanding of the increasing importance for teachers and teacher trainers to be attentive to YLs' assessment needs, and move beyond the current foundation of pedagogical and assessment skills.

7. References

Bandalos, Deborah L. 2004. "Can a Teacher-led State Assessment System Work?" *Educational Measurement: Issues and Practice* 23:33-40. doi:10.1111/j.1745-3992.2004.tb00157.x.

Bejarano, Yael, and Claire Gordon. 2009. "Consideration for Teaching and Assessing Young Learners Learning English as a Foreign Language." Retrieved from http://www.ericdigests.org/1997-4/young.htm.

Bialystok, Ellen. 2001. *Bilingualism in Development: Language, Literacy, and Cognition.* New York: Cambridge University Press.

Black, Paul, and Dylan William. 1998. "Assessment and Classroom Learning." *Assessment in Education: Principles, Policy & Practice* 5(1):7-74, doi: 10.1080/0969595980050102.

Buchanan, Elizabeth A., and Erin E. Hvizdak. 2009. "Online Survey Tools: Ethical and Methodological Concerns of Human Research Ethics Committees." *Journal of Empirical Research on Human Research Ethics* 4(2):37-48. https://doi.org/10.1525/jer.2009.4.2.37.

Butler, Katherine G. 1997. "Dynamic Assessment at the Millennium: A Transient Tutorial for Today." *Journal of Children's Communication Development* 19:43-54.

Conner, Owen. 2008. "Testing as a Catalyst for Progress and Development." *The International TEYL Journal*. Retrieved from http://www.teyl.org/article7.html.

Fulcher, Glenn. 2012. "Assessment Literacy for the Language Classroom." *Language Assessment Quarterly* 9(2):113-132.

Giannikas, Christina N. 2017. "Assessing Young Learners through Games and Enjoying it!" *TESOL Greece Magazine* 136:31-33.

Goldstein, Jeffrey. 2012. "Play in Children's Development, Health and Well-being." In *Brussels: Toy Industry of Europe* (TIE).

Green, Anthony. 2006. "Assessment Literacy for Language Teachers." In *Classroom-based Assessment in L2 Contexts*, edited by Dina Tsagari, 8-29. Newcastle upon Tyne: Cambridge Scholars Publishing.

Hasselgreen, Angela. 2005. "Assessing the Language of Young Learners." *Language Testing* 22(3):337-354.

Ioannou-Georgiou, Sophie, and Pavlos Pavlou. 2003. *Assessing Young Learners.* Oxford: Oxford University Press.

Inbar-Lourie, Ofra, and Elena Shohamy. 2009. "Assessing Young Language Learners: What is the Construct?" In *Contextualizing the Age Factor: Issues in Early Foreign Language Learning*, edited by Marianne Nikolov, 83-96. Berlin: Mouton de Gruyter.

Klopfer, Eric, Scot Osterweil, and Katie Salen. 2009. *Moving Learning Games Forward*. Education Arcade MIT.

Koh, Kim. 2011. "Improving Teachers' Assessment Literacy through Professional Development." *Teaching Education* 22(3):255-276.

Leung, Constant. 2014a. October. *Learning from Feedback: Conception, Reception, and Consequences.* New York: TCCRISLS Conference.

Leung, Constant. 2014b. "Classroom-based Assessment Issues for Language Teacher Education." In *The Companion to Language Assessment,* Vol. III, edited by Anthony J. Kunnan, 1510-1519. Chichester: Wiley Blackwell.

Lukin, Leslie E. 2004. "Facilitating the Development of Assessment Literacy." *Educational Measurement: Issues and Practice* 23(2): 26-32.

McKay, Penny. 2006. *Assessing Young Language Learners*. Cambridge: Cambridge University Press.

McMillan, James H. 2000. *Classroom Assessment: Principles and Practice for Effective Instruction*. Pearson Technology Group. Boston: Allyn and Bacon.

Ontario Ministry of Education and Training. 1995. The Common Curriculum: *Policies and Outcomes,* Grades 1-9. Toronto: Ontario.

Popham, James W. 2004. "Why Assessment Literacy is Professional Suicide." *Educational Leadership* 62(1):82-83.

Popham, James W. 2009. "Assessment Literacy for Teachers: Faddish or Fundamental?" *Theory into Practice* 48(1): 4-11. https://doi.org/10.1080/00405840802577536.

Prensky, Marc. 2001. "Digital Native, Digital Immigrants." *On the Horizon* 9(5):1-6. https://doi.org/10.1108/10748120110424816.

Scarino, Angela. 2017. "Culture and Language Assessment." In *Language Testing and Assessment. Encyclopedia of Language and Education* (3rd ed.), edited by Elena Shohamy, and Nancy Hornberger. Dordrecht: Springer Netherlands.

Shaaban, Kassim. 2005. "Assessment of Young Learners." *English Teaching Forum* 43(1):34-40.

Shute, Veronica J. 2006. April. "Assessments for Learning: Great Idea, but Do they Work?" Paper presented at the annual meeting of the American Educational Research Association, San Francisco: CA.

Smith, Karl. 1996. "Assessing and Testing Young Learners: Can We? Should We?" *In Papers from a Symposium of the Research, Testing,* *and Young Learners Special Interest Groups*, edited by Dave Allen. IATEFL, Kent, England.

Stiggins, Richard J. 1994. *Student-centered Classroom Assessment*. Merrill, Toronto: Maxwell Macmillan Canada.

Stiggins, Richard J. 2002. "Assessment Crisis: The Absence of Assessment for Learning." *Phi Delta Kappan* 83(10):758-765.

Su, Feng-Huan. 2000. "New Goal Orientation of English Education for Young EFL Learners." In *Proceedings of the conference of ELT curriculum for young learners in East Asia*, edited by The English Department, 107-20. Taipei: National Taiwan Normal University and British Council.

Taylor, Lynda. 2009. "Developing Assessment Literacy." *Annual Review of Applied Linguistics* 29:21-36.

Torrance, Paul E. 1998. *The Torrance Tests of Creative Thinking: Norms-technical Manual Figural (Streamlined) Forms* A & B. Bensenville, IL: Scholastic Testing Service.

Torres Soriano, Verónica. 2009. "Alternative Assessment: Portfolio Assessment for Young Learners." *MEXTESOL Journal* 33(1):133-137.

Tsagari, Dina, and Christina N. Giannikas (in preparation). "The Impact of Cert-mania in English Language Learning and Teaching: the Cypriot Context".

Tsagari, Dina, and Karin Vogt. 2017. "Assessment Literacy of Foreign Language Teachers around Europe: Research, Challenges and Future Prospects." *Papers in Language Testing and Assessment* 6(1):41-63. Special Issue edited by Kathryn Hill on "Teacher Assessment Literacy in Second and Foreign Language Education". http://arts.unimelb.edu.au/ltrc/research/publications/papers/current-issue.

Tourangeau, Roger, Lance J. Rips, and Kenneth Rasinski. 2000. *The Psychology of Survey Response.* Cambridge: Cambridge University Press.

Vernon, Shelley A. 2006. "Discover how your Pupils Can Learn to Speak English Twice as Fast when they're Excited about Learning. Using Hundreds of Fun English Games. Teaching English Games, Learning is Fun!" https://www.teachingenglishgames.com/esl-classroom-games.

Vogt, Karin, and Dina Tsagari. 2014. "Assessment Literacy of Foreign Language Teachers: Findings of a European Study." *Language Assessment Quarterly* 11(4):374-402. doi: 10.1080/15434303.2014.960046.

Wiggins, Grant. 1989. "A True Test: Toward More Authentic and Equitable Assessment." *Phi Delta Kappan* 92(7):81-93. https://doi.org/10.1177/003172171109200721.

Wolf, Dennie, Janet Bixby, John Glenn, and Howard Gardner. 1991. “Chapter 2: To Use Their Minds Well: Investigating New Forms of Student Assessment.” *Review of Research in Education* 17(1):31-74. doi:10.3102/0091732X017001031.

Yolageldili, Gülin, and Arda Arikan. 2011. “Effectiveness of Using Games in Teaching Grammar to Young Learners.” *Elementary Education Online* 10(1):219-229

CHAPTER SIX

WILL A BOOM LEAD TO A BLOOM? OR HOW TO SECURE A LAUNCH OF LANGUAGE ASSESSMENT LITERACY IN UKRAINE

OLGA KVASOVA

The essential role that language testing and assessment (LTA) plays as part of FL instruction has prompted the recent introduction of LTA courses on different levels of pre-service teacher training in Ukraine. However, the implementation of this long-awaited initiative may not appear successful due to the lack of expertise in the design of LTA courses and materials in the country. The researchers with such expertise feel a responsibility to provide the teaching staff with frameworks for developing courses as well as guidance in aligning these frameworks to particular educational contexts.

The chapter focuses on three LTA courses developed by the author and delivered in a classical university at undergraduate and graduate levels. Sharing a common stable knowledge base, the courses vary significantly in accordance with the learning agenda in each particular case. The author discusses the course components, the methodology of course delivery including modes and formats of instruction and techniques of training and assessing the LTA knowledge and skills of pre-service teachers. The findings of the study may be of use to teachers and institutions committed to making assessment literacy sustainable in diverse national and international educational contexts.

1. Introduction

The implementation of ECTS as well as the ever-increasing international perspective of Ukrainian (UA) higher education have drawn stakeholders' attention to FL teacher training in the use of L2 tests as tools to ensure quality language education. Although local experts have sufficiently developed theoretical and practical foundations for language assessment, the introduction of formal in-service teacher training in language testing and assessment (LTA) is still pending. Recent years, however, have witnessed an upsurge in stand-alone pre-service teacher training courses, the number of which compares with almost none a decade ago.

The long-sought government decision to introduce LTA entails multiple challenges for the university teachers who are to deliver these courses. The teachers who embark on the LTA curriculum and materials design without enhanced language assessment literacy (LAL), hope to gain sufficient expertise in LTA on the job. On the other hand, the teachers who have participated in the LAL training events realize the instructor's responsibility and reject taking up a course. At teacher development events, those teachers indicate they would develop a context-specific course in the case they received specialized training in curriculum and materials design or at least had at their disposal a framework they could build their courses on. Both situations put at risk the LAL launch in UA pre-service teacher education and may result in one more assessment illiterate generation of FL teachers. Finally, there are the teachers who have been initially concerned with the development of in-service teachers' assessment literacy, such as the author of this chapter. Having extensive experience in training university teachers in LTA, she has been engaged in pre-service teacher training introduced into three educational levels–from PhD to undergraduate students–for two academic years. This experience laid the basis for this small-scale mixed-method study whose main objective is to collect and interpret evidence which can be further used in configuring pre-service teacher training courses.

A few comments should be made about the differences between UA higher education and European or American higher education. UA students are generally overloaded with study work including contact classes (up to fifteen 80-minute classes per week). The majority of courses are compulsory and only recently have elective courses been introduced in UA universities. A bachelor's course lasts four years, whereas a master's course lasts two years. It should be specified that a master's degree is not a scholarly degree but an educational qualification in Ukraine. Since a BA is erroneously considered "incomplete" higher education, the majority of bachelor degree holders seek to enroll for a master's programme immediately upon completing the bachelor's course. Such country-specific education is not exactly in tune with educational practices worldwide. As a result, instead of practitioners with some years of teaching experience, and clear about their professional goals and scientific interests, master's degree seekers in the UA context are student teachers rather than researchers. Similarly, a PhD course is also considered a pre-service course from the perspective of this chapter.

The comparison of three courses aimed at diverse target audiences, implemented by one trainer in one particular institution is expected to shed some light on the relationship between assessment knowledge, principles and skills within each course content. It is also expected to establish the ways to connect curriculum, instruction and learning with a view to meeting social demand in assessment-capable teachers. The empirical data of the study will allow the identification of the most relevant curricular components which will enable training course configuration for specific target audiences.

2. Overview of Theoretical Grounds for Curriculum Development

The issues of teachers' language assessment literacy (LAL) were initially addressed by applied linguists in the 1990s (i.e., Brown 1996; Johnson et al. 1999, cited in Hatipoğlu 2010) and remain debated today. In fact, the concept of LAL is scrutinized by every researcher engaged in designing LTA courses, as the concept is crucial for defining the content in the curricula in various settings.

The vision of the LAL construct has changed over the years. In 2008, Brown and Bailey analyzed LTA courses in order to compare their content with similar data obtained in 1996. They argued that the majority of courses covered the so-called *stable knowledge base,* "that is evolving and expanding rather than shifting radically" (Bailey and Brown 2008, 371). The core topics of LTA training included the principles and practice of item writing, validity and reliability issues, item analysis, etc., which help teachers learn to write language tests properly.

In a recent study, Stabler-Havener (2018) provided a comprehensive analysis of the key LAL models developed by Brindley (2001), Inbar-Lourie (2008), Davies (2008), Fulcher (2012), Pill and Harding (2013) and Taylor (2013). She points to general similarities among the models but maintains that "each model tends to articulate LAL competencies differently", which may create challenges for stakeholders in utilizing the models. Moreover, "another problem with the models is that although most of them provide general definitions of the competencies and some provide specific examples, none of them operationalize their model with detailed, descriptive specifications as to what a language assessment literate person needs to know and do in order to be considered language assessment literate" (Stabler-Havener 2018, 9). She further notes that such lack of specificity is an asset for the course designers who are assessment literate, as it allows them "to select, adapt, and operationalize a model most appropriate for their context". In the opposite case, testing professionals should endeavour to equip non-expert course providers with articulated objectives and detailed specifications as well as guidelines on how to adapt the curricula to specific academic contexts (Jeong 2013, cited in Stabler-Havener 2018). In this line of argument, bridging the gap between theory and practice–conceptualization of LAL and its operationalization for instruction purposes in particular situations–is obviously viewed as an objective of LTA curricula and course design.

The passing decade has marked the researchers' essential contribution to providing frameworks for pre- and in-service teacher education in LTA. The courses developed within two Erasmus projects, PROset (Green 2014) and TALE (Tsagari et al. 2018), have promoted the transition of LTA courses from being "marginal, too short, mostly elective and with the scope too narrow for future specialists", as Taylor characterized them in 2009, to well-grounded and well-structured courses for face-to-face and online provision. Both courses cover a similar *stable knowledge base* specified according to the needs of classroom-based assessment surveyed by Vogt and Tsagari (2014). Classroom teachers are not engaged in sophisticated statistical analyses as are the developers of standardized tests. For teachers, Taylor (2013) argues, it is of primary importance to know the language pedagogy along with technical testing skills, local practices and sociocultural values as well as personal beliefs and values. The teachers' LAL profile is complemented by knowledge of theory (principles and concepts) and skills in scoring and decision-making. Student teachers' command of language pedagogy and language teaching approaches is also viewed as an important prerequisite for the effective translation of the general assessment literacy into language related know-how (Inbar-Lourie 2008). In line with the above, an LTA course has its roots in language pedagogy, whereas LAL itself is individualized, i.e. the product of knowledge, experience, and perceptions that the language teacher brings to the teaching and assessment practice (Scarino 2013, cited in Inbar-Lourie 2013). After defining the course construct, another question should be addressed, i.e. *in which ways* should the theory, practice and principles of LTA be communicated to trainees. In line with a learner-centred approach, such as the one enabling trainees' agency, these ways should be consistent with the students' future professional needs, learning motivation and mental maturity. The language assessment culture adopted in the community should also be taken into account when choosing instruction scenarios, especially in respect of the compatibility of assessment practices with institutional or individual ideologies, social expectations, attitudes and values (Inbar-Lourie, 2008).

Apart from the two major courses on LTA mentioned above, a great number of researchers report in detail about their courses thus providing insights from diverse educational and cultural perspectives (Kleinsasser 2005, O'Loughlin 2006, Volante and Fazio 2007, Jin 2010, Hatipoğlu 2010, Jones 2014, Lam 2015, Beziat and Coleman 2015, Solnyshkina et al. 2016, Hildén and Fröjdendahl 2018). Nevertheless, it is generally conceded that there is still plenty of room for further in-depth investigations. The current study presents yet another attempt to empirically explore curricula design and course delivery in terms of establishing instructional effectiveness and learners' perceptions of it. The implications for course provision are expected to be equally meaningful for UA teacher education as well as for the contexts where LAL in pre-service teacher training is still in its very infancy.

3. Research Design and Methodology

This study was originally conceived as small-scale for a number of reasons. It was conducted by the author of this chapter, who undertook three distinct roles as: a) a curriculum and materials designer; b) an instructor delivering the LTA courses as part of a regular lecturer's workload; and c) a researcher who developed the methodology of the study, collected and interpreted evidence, and formulated the implications for teaching LTA to various cohorts of student teachers.

Participants in the study were: two PhD students, one of whom was working on a thesis in FL teaching methodology and the other was working in the methodology of teaching foreign literature; 18 master's students majoring in TEFL; and 21 bachelor's students undertaking a training course in teaching L1 (Ukrainian) and L2 (English). Accordingly, three stand-alone courses on LTA were developed and delivered during four terms, with each course lasting 15 weeks. All courses were delivered in English.

The investigation intended to resolve the following *research questions*:

a) What content should be covered by three stand-alone LTA courses delivered at different levels in one institution? Which content should be core in each case?
b) What course delivery scenarios, including modes of instruction, training techniques and methods of assessing trainees' progress, will be relevant to each of the courses and maximize learning?

The methodology of investigation includes both qualitative methods, such as questionnaires, interviews, and observation, and quantitative ones, such as a test of LTA knowledge.

Two different *questionnaires* were administered to two of the three cohorts of trainees–at master's and bachelor levels; for bachelors majoring in Ukrainian the questionnaire was compiled in their mother tongue. Certain sections of both questionnaires had identical aims: the introductory one intended to get information about the trainees' prior academic history (educational subjects and practice that the students had undertaken) and the professional paths they were to take in the future; and the concluding one intended to elicit an evaluation of the course content and delivery, and provide comments and suggestions for improvements. One part of the questionnaire was of the Likert type inviting the respondents to "totally agree", "quite agree" and "disagree" with the statements evaluating the course components and methods of delivery. Another part of the questionnaire aimed to ascertain the learner-centredness of the course by ranking course components and methods of delivery from most significant/appropriate to least significant/appropriate. The questionnaire also contained sections to be completed by respondents willing to share their personal impressions of the training experience.

Another tool used in this study was a test developed and used in a study of university teachers' assessment literacy by Kvasova (2016) to measure the effectiveness of training. The original test consisted of selected response items focused on knowledge of the major principles of classroom test development (testing cycles, test usefulness, testing techniques, item writing). In this research, the PhD and MA level students were offered a 50-item version of the test whereas the bachelor level trainees did a shorter, adapted version of the test (20 items).

Classroom observation data and comments on formative assessment are provided in the section describing the findings.

4. Findings and Discussion

This section will present the data regarding the content, resources, and methodology of delivery including the mode of instruction, teaching techniques and trainees' evaluation of the course, as well as the efficacy of the training within three stand-alone courses on LTA.

4.1. A course for PhD students

Since the number of PhD students in teaching methodology is quite limited in the classical university where the researcher works, the course on LTA was offered as elective. So far it has been taught to two PhD candidates in one academic year.

The curriculum for PhD students was based on a previously practised curriculum for in-service teacher training (ibid. Kvasova) which had been updated as the author had been gaining theoretical and empirical expertise. The course was intended for PhD candidates working in the domain of foreign language teaching methodology since all PhD theses on teaching methods presuppose measuring learners' efficacy in attaining objectives of instruction. Given the academic and research background of the PhD candidates, the course was originally aimed to reveal the prior LTA knowledge and skills and bridge possible gaps in theory. It was also intended to equip trainees with skills in test development and data interpretation, thus ensuring the adequate use of language tests as research tools. Furthermore, it was agreed to tailor the instruction to the candidates' research needs.

Both candidates were recommended to take the free online courses developed by the Erasmus+ TALE project, at least two out of eight–the introductory one and the one related to their research. A bibliography list included the most significant research articles and monographs accessible in Ukrainian contexts along with the textbook on LTA written in Ukrainian by the author of this chapter. One of the trainees received instruction via face-to-face tutorials and the other via distance learning, online tutorials. The tutorials were conducted as a discussion of the theoretical issues and practical assignments were delivered by the PhD candidates independently. In the case of one trainee, an additional focus of the tutorials was on the specific methods of investigation applied in her PhD research. The other trainee received additional supervision through recommendations to study particular papers on her research topic, consultations and suggestions to improve the test tasks designed.

The motivation for obtaining the high learning outcomes of this course was supposed to be high, given the candidates' education, their teaching experience of 2-7 years, and a clearly articulated need to gather and accurately interpret the data verifying the effectiveness of their teaching methodology. However, on the exit test, both trainees scored slightly higher than 70%, i.e. 72% and 74% respectively, which was disappointing to the candidates but quite comparable with the scores obtained from serving university teachers who participated in the earlier research. In the case of the practitioners an explanation could be overestimation of their assessment knowledge and skills and an underrating of special LTA training. In the case of the PhD students, based on the observations made during the course provision, certain alterations should be introduced in the course content, delivery and assessment of outcomes. To be consistent with the requirements of the academic degree sought, the content should contain a broader scope of themes and more sophisticated concepts, tutorials, in their majority, should be replaced by seminars during which prospective PhD holders should report on their knowledge about the use of language tests as tools of pedagogical investigations. Academic essays revealing the application of these tools within a framework of a particular research study should be assessed against the criteria developed by the instructor, with the trainees' possible involvement in the process. The course should focus on the trainees' ownership of the developed knowledge and skills, and enhance responsibility for the quality of their research.

4.2. A course for master's students

The compulsory course was introduced following its initial implementation as *elective* to master's students majoring in TEFL (n=8), which can be considered as course piloting. The curriculum encompassed the *stable knowledge base* and the instruction was delivered through an alternation of lectures and workshops. The eight students scored on the exit testing 70% on average with one of the trainees achieving 96%. Although the outcomes of the course acquisition by these trainees are not examined in this study, the delivery of the course offered some meaningful insights about the students' involvement and achievements. It appeared that the trainees of the younger generation were test-wiser than the serving teachers and PhD candidates possibly because of their having taken the independent school-leaving exam or commercial tests of proficiency in English. Half of the trainees provided quite grounded responses to theory-related questions and coped with practical assignments (writing items and tasks) with relative ease. Unlike PhD candidates, the outcome of the course for MA students was encouraging: the curriculum proved to meet the needs and mental/cognitive maturity of the 22-23-year-olds and seemed to be almost precisely tailored to them. The course implementation and outcomes were highly evaluated by the educational administrators, which enabled the change of the course's status from elective to compulsory in the following academic year.

The compulsory course for master's students majoring in TEFL was taken by 18 trainees. Based on the experience of the "pilot" course, the curriculum underwent minor alterations, mostly in respect of the theoretical material and practical tasks on assessing productive skills. The overall material became more comprehensive, more elaborate and was accompanied with video recorded materials and undergraduates' original writings to mark and grade to enable the development of rating scales in teams. The trainees were provided with a list of open access English language sources available, as well as the trainer's presentations and handout materials. The demand for material was growing as the exit test was approaching. In the mode of instruction there was a shift towards greater balance between theoretical input and hands-on experience, so the lecture format was totally rejected and replaced with workshops. Each workshop enabled the trainees to complete some practice-related tasks. This was teamwork that focused on item/task design, trialling, feedback provision to other teams and further modification of tasks either in the classroom or outside it.

Observation data from the trainees' attendance and participation in constructing tools of assessment indicate the following: the attendance of face-to-face classes was marked with varied frequency (from 20 to 100% of classes) and relatively efficient participation. Similarly, the average number of individual practical assignments done by the trainees and submitted for teacher evaluation ranged from the lowest, 2 to the maximum of 5. With respect to the quality of the student-constructed test items/tasks, it was generally quite high given that the item writers were beginners. If benchmarked against the in-service teacher-constructed test tasks which were examined with the help of expert judgements within the prior research, the student-constructed items/tasks definitely exceeded expectations. This evaluation does not exclude a share of subjectivity as the opinion of the instructor was not reinforced by other experts' judgements. However, while discussing the tried-out tasks in the classroom, the students made very relevant points about their quality and voiced quite mature suggestions for improvement. The exit test also demonstrated a quite high level of material acquisition.

% of correct answers	Number of testees	% of testees
96	2	11
95	1	5.5
94	5	27.5
92	3	17.5
90	1	5.5
86	1	5.5
82	1	5.5
80	1	5.5
74	1	5.5
72	1	5.5
70	1	5.5

Table 1: Scores on the exit test (master's students)

According to Table 1, the overwhelming majority of trainees, i.e. 12, which totals 67% of all trainees, have attained scores equal to or higher than 90%, which equates to grade "A" according to the ECTS scale. Such results were unexpectedly high compared with the results of the PhD candidates and in-service teachers. After the test had been administered, the testees were interviewed on the difficulty of the tasks included in the test. Six of them indicated that the test was overall quite easy with more difficult tasks appearing towards the end. Another six claimed the test was of average difficulty and the remaining six admitted it was difficult for them, especially the final tasks. Apparently, the latter evaluations were voiced by the interviewees (n=6) who attained the six lowest scores and were evidence of their worse grasp of the material.

At the end of the course, the trainees were surveyed to evaluate the curriculum and the methods of course delivery. There were 16 respondents and 2 questionnaires were invalid (completed inaccurately). The *questionnaire* helped to elicit the following information about the *course content relevance* to MA students' career prospects. As it appeared, out of sixteen respondents, eight were sure they would work as teachers (Ts), six preferred a career in all possible domains but education (NTs), and two respondents were not certain about their vocational orientation and could not respond to this question.

Speaking about the *background in educational sciences*, all respondents but three indicated that they had previously taken a course in general pedagogy, and all sixteen had had training in foreign language teaching methodology and practice in secondary schools. Six respondents were currently engaged as private tutors, another three were employed as tutors of private FL courses, and one of the respondents mentioned she was preparing private students for commercial examinations. Also, the respondents were asked to indicate if *the course components* taught were relevant to their needs and interests and if they should be included in the curriculum. The top scores for relevance were received by such components as "ensuring reliability of tests" and "assessment of writing skills" (16 of 16), closely followed by "test purposes", "test specifications", "practicality", and "impact", as well as "assessment of speaking skills" (15 of 16). The issue of "test security" (12 of 16) had the lowest relevance.

The following question aimed to determine the *ranking of course components* with respect to the needs of the future profession. Here the rating of course components was processed across the two groups of respondents–Ts and NTs (14 respondents). Figure 1 represents a comparison of ranking the curriculum components by Ts and NTs.

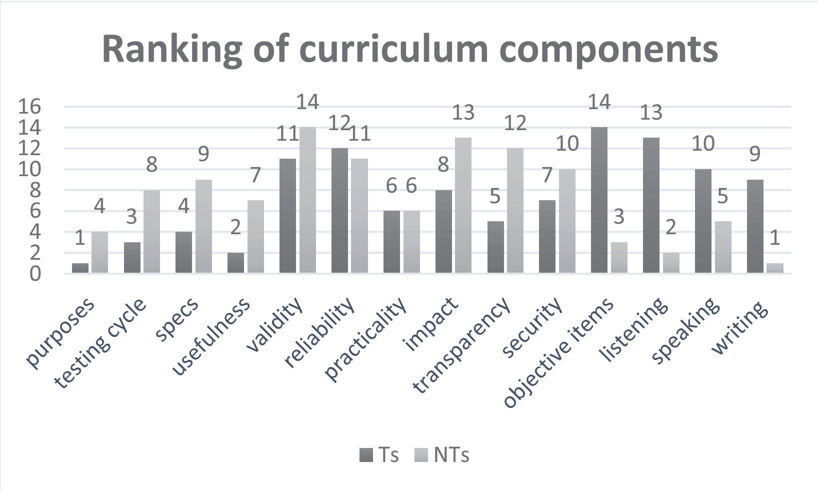


Figure 1: Comparison of curriculum component rankings "Teachers vs. Non-teachers"

The numbers show that for Ts, learning how to "write objective test items for reading", "mapping texts for listening test task development", "assessment of speaking" and "assessment of writing", that is, all the issues concerned with the practical application of LTA knowledge, are ranked higher than by their counterparts. "Reliability" and "validity" ranked high as the major cornerstones that ensure quality of assessment. The lowest positions in the ranking were taken by theory-related components (specifications, testing cycle, test usefulness and test purposes). In the follow-up interviews of five prospective teachers, however, they clarified that they had not underrated the significance of these concepts but had taken them for granted and would bear them in mind while preparing tests.

For NTs, conversely, the practice of developing items and tasks is not as valuable as for Ts. They rank highest "validity", "transparency", "reliability", "security", and "practicality" along with other theory-related components (usefulness, specifications, and testing cycle). This ranking gives NTs credit, although the importance they attribute to these issues seems to display their concern about being fairly assessed as test takers, in the first place. Among the assessment skills, which were ranked the lowest by this category of respondents, the most notable interest was displayed towards the "assessment of speaking", which is still controversial in Ukrainian language pedagogy.

The most typical respondent comments regarding the content of the curriculum were:

> *"... the content of the course was full of important things such as assessing reading, writing, listening and speaking but I believe we haven't had enough time to cover all the materials in detail. The course is important enough to last for two terms since it can empower prospective teachers to become fair assessors."*

> *"All the aspects of the course were crucial for future teachers and test writers".*

> *"I was mostly interested in such issues as practicality and validity. What we were presented with was very helpful".*

> *"The course content helped me to understand strategies to cope with some test tasks, how to prepare materials for assessment and, on the whole, how to organize my work."*

Two of the respondents pointed out, however, that *"the content was interesting but, in some cases, obvious*" and *"was not new*" to the respondents.

Information about the evaluation of course organization/delivery was elicited by means of marking the statements with "fully agree", "quite agree" and "disagree". As seen in the diagram, the predominant number of responses lie within the ranges, "fully agree" and "quite agree".

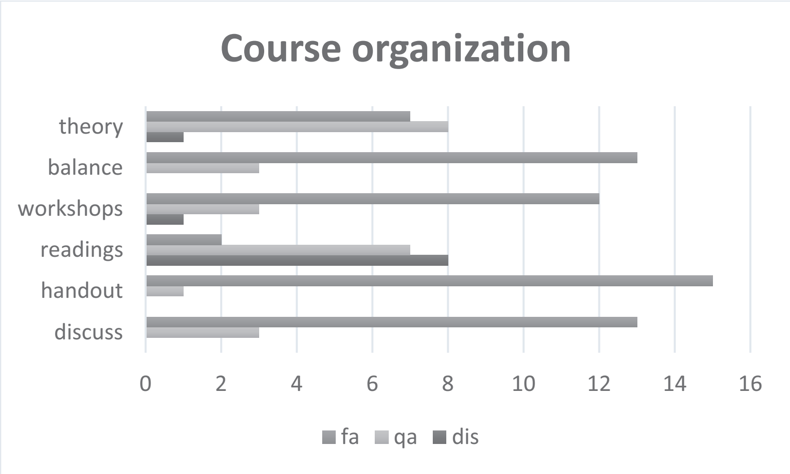


Figure 2: Ranking of Course Organization Components

In fact, disagreement was expressed regarding half of the evaluative statements (three out of six). One of the respondents did not agree that *“there was too much theoretical input*” and another that workshops seemed *“more effective than lectures*”. Surprisingly, this is the third incidence of “disagreement” expressed by eight respondents out of sixteen: they did not find it necessary to assign additional readings on the themes studied. Perhaps this was due to their difficulty in obtaining recommended sources, or their overall workload. At least, almost unanimously the respondents were pleased with the handouts used to complement presentations and the materials provided for practical assignments. There was predominant agreement regarding the appropriateness of the balance between theoretical input and practical tasks, as well as the opportunity to discuss assessment issues with fellow students and the trainer.

The course was very well received by all categories of respondents who indicated that the course had met their expectations of being “comprehensive” (13) and “up-to-date” (16). 13 respondents out of 16 noted that the course on LTA should be compulsory and not elective within an MA course, and “allotted more contact hours to enable teamwork and face-to-face guidance of the trainer”. 14 of the respondents stressed that the course “should have been taught earlier” within the undergraduate programme, immediately following practice in schools. This, as students commented, would be more motivating for them and effective in terms of raising awareness of the practical application of LTA.

Other comments on the course related to the workshop format of course delivery which is still not established in the Ukrainian education system and therefore was appreciated as innovative. The respondents gave credit to the clear presentation of the complex theoretical concepts of LTA and the use of multiple tasks to promote thinking about them. At the same time, the majority of respondents suggested that more time should be dedicated to training in item/task writing and the use of rating scales.

The above findings explicated the MA students' willingness and capacity to embrace both core and peripheral ideas in the area of language assessment as the course content seemed to perfectly meet their professional aspirations and/or practical needs. In the exit test, this cohort of trainees outperformed the serving teachers and PhD candidates and in this respect the data of this study echo the data obtained by researchers of LTA training in other educational contexts (e.g. Alkharusi et al. 2011, cited in Beziat and Coleman 2015). Similar to the evidence obtained by Volante and Fazio (2007), this study ascertains the desire and readiness of students to gain greater practical knowledge about assessment.

4.3. A course for bachelor students

The LTA course at the undergraduate level was implemented immediately after the courses on general and language pedagogy and the practicum at secondary school. This suggested that LTA knowledge and skills should be built on a recently gained theoretical knowledge/ground and reinforced by experience of teaching. The former was expected to strengthen interdisciplinary ties enabling better comprehension of domain-related concepts and processes, and the latter to foster motivation towards developing career-relevant skills.

An essential point to make is that the students of this cohort have a major in teaching Ukrainian and TEFL is their minor. Given the lower level of English proficiency of these students, the course *de facto* possessed characteristics of content and language integrated learning and therefore needed certain reprofiling. The course was delivered to 21 second-year students, aged 18. Notwithstanding their immature age, the trainees themselves were fairly test-wise having prior experience in taking standardized tests and teacher-constructed tests at the university. The course content was left basically the same but the course delivery and teaching materials were essentially tailored to the students' age, background knowledge and English proficiency.

The course was implemented as a series of workshops based on the generic principles of this format, e.g. a lead-in to elicit students' perceptions and prior experience on the theme of the class, presentations of the new material supported with numerous handouts, thus facilitating exploratory activities, and encouragement of teamwork at every moment of the class. Still, the trainer's speech had to be adapted to the learners' abilities by using simpler structures and more frequent lexis, resorting to an explanation of meta-language and delivery at a slower pace. Thus, the new concepts were immediately translated into the mother tongue. In fact, students preferred to have the presentations in advance to get prepared to comprehend them properly in the classroom. Also, at their request, short glossaries were also prepared for almost each class. The continuous observation of course implementation revealed that the multiple and diverse handouts played an essential role in maximizing instruction by scaffolding learning activities.

The questionnaires completed by 20 students elicited that they were "*excited to have a course delivered in English because additional language practice is always useful*", and they enjoyed learning interactively and collaboratively, valuing the opportunity to discuss issues and do tasks in the classroom. Concerning the use of English as a medium of instruction, it was supported by 90% of respondents. For half of the respondents, learning subject content through the medium of English was not difficult whereas the other half indicated it was quite challenging. Those with challenges noted, however, that "*it was not a big obstacle since the teacher explained everything clearly and translated the new terms"*. The hope to improve their English proficiency while doing the course was expressed by all but one of the respondents. The students' appreciative responses implied that the trainer had done a lot of work to ensure scaffolding, which allowed the majority of students to pass the major linguistic challenges unnoticed and feel satisfied with self-efficacy.

As for the conceptual content of the course, 50% of respondents (n=10) who admitted their own insufficient language proficiency were divided in evaluating their achievements in the subject: it was equally "difficult" for three, "quite difficult" for two of them while the remaining 5 trainees claimed they could manage subject acquisition perfectly well. When it came to distinguishing which means helped students acquire the content in English, all respondents noted the primary importance of handouts and the availability of a test book written in their mother tongue. The overwhelming majority of respondents maintained that their better acquisition of the subject was promoted by collaborative work and team discussions. None of the respondents indicated they would want to return to the lecture format of instruction. It was also interesting to learn which factors (background knowledge, prior experiences, and job prospects) mostly facilitated the trainees' motivation toward excelling in the course. It appeared that 70% (totally) and 20% (partially) credited their efficient course acquisition to their own test-taking experience, whereas 10% said it had no impact at all. 90% of respondents noted that they were motivated by future professional needs and 90% attributed this motivation to the stimulating course content and the fashion of its delivery.

In the comments made by the respondents there were a few insightful suggestions: several comments noted that "*the course should be allotted more time on the whole and more face-to-face hours, in particular",* and *"the learners want to work in teams interacting with each other, write items together and help each other to improve them"*. Many respondents expressed the desire to dedicate more classroom time to *"constructing tasks"* and one respondent came up with the idea of providing the trainees with "*more frequent feedback as formative assessment"*. The exit test yielded an average 82% of correct answers with the percentages ranging from 70 to 94. Although quite indicative of the effectiveness of the course, the data cannot be considered a reliable source of information about the testees' achievements. Obviously, it needs major revision and adaptation.

In conclusion, the majority of the students were highly motivated to excel in the subject. Their attendance and preparedness of tasks for trialling and feedback testified to their involvement in the course in terms of both their interest towards more theoretical issues (validity, reliability, practicality) and the active application of rules of item and task writing. What impeded them in doing practical tasks was the perspective they still adopted–not that of a test writer, but that of a test taker. This obvious lack of confidence was worrying the instructor during several of the initial classes, after which the "test taker's complex" was gradually overcome by a greater part of the trainees thus dissipating the doubts of course inappropriacy to such an immature cohort of students.

5. Conclusion

The implementation of the three courses on LTA discussed in the chapter allows the voicing of the following implications including the prospects of further research in the domain.

The core components of the course are related to the *stable knowledge base* and are complemented by the most relevant classroom testing skills. They were established via profiling LAL for particular groups of prospective teachers–undergraduate and graduate students. The scope of theoretical input was determined regarding the perceived professional needs, educational background and teaching experience of each of the target audiences. Unfortunately, the duration of the course does not allow the inclusion of some material, such as alternative assessment within the core course components, which is viewed as a shortcoming. However, the spread of contact hours shows that, following the respondents' suggestions, a (relatively) sufficient time is allotted to the development of the so-called technical skills.

No/week	Title	Theory	Practice	Assessment
1/1	ABCs of LTA	Purposes of tests; types of tests; parameters of test usefulness; principles of test development; scoring; feedback.	Development of test task. Test administration.	-
2/2-3	Task types	Components of tasks; the rubrics; test formats; types of response (selected response, limited production and extensive production).	Development of test tasks with different response types. Developing tasks to check grammar skills (alternative choice, gap-filling, paraphrase).	Formative: questioning, doing quizzes, presenting team-made tasks for discussion and feedback.
3/4-6	Assessment of reading skills	Types and skills of reading; CEFR descriptors for reading; test specifications; reading test tasks' preparation procedure.	Developing tasks to test reading (multiple-choice questions, true/false).	(as above)

4/7-8	Assessment of listening skills	Listening test tasks' preparation procedure (test mapping technique).	Developing various tasks based on one listening test (multiple-choice questions, true/false, summary completion, short-answer questions).	(as above)
5/9-11	Assessment of speaking skills	Speaking part of acclaimed examinations (video presentation); CEFR and speaking test specifications; types of rating scales; tasks for classroom assessment of oral production.	Evaluating a recorded speaking performance against an analytic rating scale. Developing tasks for classroom assessment of oral production and analytic scales.	(as above)
6/12-14	Assessment of writing skills	CEFR and writing test specifications; types of rating scales; tasks for classroom assessment of written production.	Grading and scoring of sample papers. Developing tasks for classroom assessment of written production and analytic scales.	(as above)
				Summative test; evaluation of five individually designed test tasks

Table 2: A sample framework curriculum

It is proposed that the core components are identical in each curriculum but tailored to the cognitive abilities/mental maturity of the trainees, depending on their age and learning experience, as well as the learning motivation and ambitions connected with the trainees' career prospects, their attitude to teamwork and attitudes to being evaluated by peers and/or the trainer. Similarly, learner abilities should be central in determining the methods and techniques of developing assessment skills, regulating the amount of practical assignments and the mode of their presentation for discussion. The younger the trainees, the more flexible they are in adopting a new format, and new interactive and collaborative styles of instruction. In the case of content and language integrated learning, learning needs to be sufficiently scaffolded but remain stimulating to trainees.

Unquestionably, formative assessment is to be continuously realized through teachers' observation of trainees' engagement in learning activities, their participation in teams and individual task preparation. There is also a lot of room for self- and peer-assessment to be integrated in the course, e.g. not only in trialling and providing feedback on peer-constructed tasks but maybe involving trainees in the development of formative assessment materials for the course. As for the summative test, it should be attuned to each cohort of learners.

The availability of resources appears essential too. Ideally, a handbook for the teachers who endeavour to teach LTA courses, as well as an updated and photocopiable resource pack for students, need to be urgently compiled. These resources aim to provide curricula/materials designers and course providers with a sound foundation for self-education in LTA and to ensure practicality in course delivery. Moreover, the teachers engaged in training students in LTA should be initially trained themselves in order to feel confident and highly professional in the domain. They need to develop the profile that comprises core LTA knowledge and technical skills, and interweave these with knowledge of local practices, societal and personal values and beliefs, and further nurture them in their trainees.

The current study provided empirically based insights into determining LTA course content and its tailoring to specific educational situations within one institution. The findings and implications may promote a sustainably high level of teachers' assessment literacy across generations and countries.

6. References

Brown, James Dean, and Kathleen M. Bailey. 2008. "Language Testing Courses: What are they in 2007?" *Language Testing* 25(3):349-83.

Green, Anthony. 2014. "PROset Summary" available online: http://proset-tempus.net/pluginfile.php/2/course/section/2/ProSET%20Summary.pdf (last accessed 15 August 2014).

Hatipoğlu, Çiler. 2010. "Summative Evolution of an Undergraduate 'English Language Testing and Evaluation' Course by Future English Language Teachers." *English Language Teacher Education and Development* (ELTED), 13 (Winter 2010): 40-51.

Hildén, Raili, and Birgitta Fröjdendahl. 2018. "The Dawn of Assessment Literacy–Exploring the Conceptions of Finnish Student Teachers in Foreign Languages." *Apples: Journal of Applied Language Studies* 12(1):1-24. Retrieved from doi:10.17011/apples/urn.201802201542.

Jin, Yan. 2010. "The Place of Language Testing and Assessment in the Professional Preparation of Foreign Language Teachers in China." *Language Testing* 27(4):555-584.

Inbar-Lourie, Ofra. 2008. "Constructing a Language Assessment Knowledge Base: A Focus on Language Assessment Courses." *Language Testing* 25(3):385-402.

Inbar-Lourie, Ofra. 2013."Language Assessment Literacy: What are the Ingredients?" Paper presented at the 4th CBLA SIG Symposium, University of Cyprus, November, 1-2.

Kvasova, Olga. 2016. "A Case of Training University Teachers in Developing and Validating Classroom Reading Test Tasks." In *Classroom-based Assessment in L2 Contexts*, edited by Dina Tsagari, 54-74. Cambridge Research Publishing.

Lam, Ricky. 2015. "Language Assessment Training in Hong Kong: Implications for Language Assessment Literacy." *Language Testing* 32(2):169-197. Accessed December 21, 2018.
DOI: 10.1177/0265532214554321.

Solnyshkina, Marina I., Elena N. Solovova, Elena V. Harkova, and Aleksander S. Kiselnikov. 2016. "Language Assessment Course: Structure, Delivery and Learning Outcomes." *International Journal of Environmental and Science Education* 11(6):1223-1229.
DOI: 10.12973/ijese.2016.392a.

Stabler-Havener, Michelle L. 2018. "Defining, Conceptualizing, Problematizing, and Assessing Language Teacher Assessment Literacy." Teachers College, Columbia University Working Papers in Applied Linguistics and TESOL 18(1):1-22.

Taylor, Lynda. 2009. "Developing Assessment Literacy." *Annual Review of Applied Linguistics* 29:21-36. https://doi. org/10.1017/S0267190509090035.

—. 2013. "Communicating the Theory, Practice and Principles of Language Testing to Test Stakeholders: Some Reflections." *Language Testing* 30(3):403-412. https://doi. org/10.1177/0265532213480338.

Tsagari, Dina, Karin Vogt, Veronica Froelich, Ildiko Csépes, Adrienn Fekete, Anthony Green, Liz Hamp-Lyons, Nicos Sifakis, and Stefania Kordia. 2018. *Handbook of Assessment for Language Teachers*. Retrieved from: http://taleproject.eu/. ISBN 978-9925-7399-1-2 (digital).

Vogt, Karin, and Dina Tsagari. 2014. "Assessment Literacy of Foreign Language Teachers: Findings of a European Study." *Language Assessment Quarterly* 11(4):374-402. DOI: 10. 1080/15434303.2014.960046,

Volante, Louis, and Xavier Fazio. 2007. "Exploring Teacher Candidates' Assessment Literacy: Implications for Teacher Education Reform and Professional Development." *Canadian Journal of Education* 30(3):749-770. https://doi. org/10.2307/20466661.

CHAPTER SEVEN

NATIONAL LITERACY ASSESSMENT AND LANGUAGE ASSESSMENT LITERACY

YIASEMINA KARAGIORGI
AND ALEXANDRA PETRIDOU

Various national testing processes–often involving standardized tests–are implemented widely not only to monitor the quality of educational systems, but also for summative and formative student assessment. Hence, beyond their daily engagement with classroom assessment, language teachers need to familiarize themselves with such external testing. This study focuses on the Programme for Functional Literacy (PfL), implemented in Cyprus to identify students “at risk” at two stages of compulsory education–Years 3 and 6–across all public primary schools of the country, so as to effectively support these children within the educational system. The chapter argues about the potential of this formative longitudinal national assessment programme to enhance certain aspects of teachers’ language assessment literacy. Key methodological characteristics of the programme are presented, while the role of teachers as key stakeholders is thoroughly discussed, especially with regard to the use of data to support teaching and learning.

1. Introduction

The role of data in public education has never been greater (Kahl et al. 2013). As Popham (2009) argues, teachers exist in environments where test-elicited evidence plays a prominent instructional and evaluative role. This has, subsequently, led to a phenomenal increase in the “testing and assessment responsibilities placed upon language teachers” (Fulcher 2012, 113).

Since they are not born testers (Jin 2010), language teachers around the world need to develop assessment literacy, i.e., “the ability to design, develop and critically evaluate tests and other assessment procedures, as well as the ability to monitor, evaluate, grade and score assessments on the basis of theoretical knowledge" (Vogt and Tsagari 2014, 377). That said, quite often teachers and test users have difficulty in developing, selecting and using tests and interpreting results, as they may have a limited understanding of assessment fundamentals (Malone 2013). According to Lam (2015), most language teachers remain underprepared to interpret the assessment information for improving teaching and learning. Various studies (e.g. Giraldo 2018) focus on the need for university programmes for prospective teachers and initial teacher education to address the development of assessment literacy. Others argue about the need to make assessment literacy an important component of teachers' professional development (Djoub 2007; Coombe et al. 2009). At the same time, to inform instructional decisions for student achievement, teachers should be supported in practice, to examine various types of data and think about using them (Thomas and Huffman, n.d.).

The chapter argues about the potential of an external formative assessment programme, the Programme for Functional Literacy (PfL), to enhance certain aspects of teachers' language assessment literacy. The PfL aims to identify students "at risk" at two stages of compulsory education–Years 3 and 6–across all public primary schools of the country, so as to effectively support these children within the educational system. To shed light onto how this central assessment programme can contribute towards the development of teachers' language assessment literacy by involving them in both data collection and use, this chapter presents key methodological characteristics of the PfL, in terms of key features and stages.

2. Literature Review

2.1 National Literacy Assessment

The national administration of standardized tests and centrally set examinations as a form of student assessment is widely used in Europe (European Commission 2009). National assessment, as the European Commission (2009) argues, falls within three categories. To begin with, to monitor the quality of educational systems, results are used as indicators of the quality of teaching and/or school effectiveness. Secondly, summative testing on individual achievement at the end of a particular educational stage can impact a student's educational career through certification, school choice and progression. Finally, national tests can be used as formative assessment to assist individuals' learning processes by identifying learning needs and, accordingly, adapting teaching; in fact, one-third of European countries employ such testing.

Beyond Europe, national testing is also frequently used through large-scale assessment for all students in Grades 3 to 8 and 10 in the United States (Davies and Elliott 2012). Canada employs both high-stakes testing at the individual level to impact entry to institutions and affect student progress at school and "low-stakes assessment" involving the confidentiality of participants (Gambell and Hunter 2004). In Australia, since 2008 the National Assessment Program–Literacy and Numeracy (NAPLAN) has required the testing of students in Years 3, 5, 7 and 9 in reading, writing, language conventions and numeracy (Dreher 2012, Hipwell and Klenowski 2011, Unsworth and Chan 2009). As Hardy (2015) argues, high results on these standardized literacy and numeracy tests are increasingly valued capitals, with implications about the practices appreciated in schools.

In Cyprus, systematic national testing to monitor the quality of teaching is not regularly exercised. With regard to summative student assessment, national exams–the Pancyprian Examinations–are only administered to students at the age of 18 (i.e., three years after the completion of compulsory education) for graduation and university entrance (Ministry of Education and Culture 2014). On the other hand, formative student assessment is usually the responsibility of teachers, who mainly employ a variety of assessment techniques, such as written tests, classroom observation, communication, and pupils' self-evaluation (Eurydice 2016). In the absence of universal screening tools and criteria at the national central level, the process of identifying students with low basic skills was challenged in the past.

This led to the development of the Programme for Functional Literacy (PfL) in 2007. Although the PfL was initially addressed only to Year 6 students, since 2011 tests have also been developed and administered to Year 3 students (Karagiorgi and Petridou 2019). Year 6 signals the end of primary education, while Year 3 tests skills to introduce remedial measures. The employment of screening measures at two stages is considered important as failure in measurements early enough may be attributed to students' slower maturation process (Catts et al. 2001, Demetriou and Kyriakides 2006).

2.2 Language Assessment Literacy

Within the accountability paradigm, beyond their daily engagement with classroom assessment, teachers often need to familiarize themselves with locally developed or externally mandated tests. The devolution of responsibility for assessment to teachers and the shift from assessment of learning to assessment for learning made it important to develop teachers' assessment literacy (Hill 2017). Language assessment literacy seems to address these two notions–assessment of learning and assessment for learning–emphasizing both classroom and external assessment practices (Jin 2010).

The increased demand for and use of assessment data by a more diverse group of stakeholders than before, has led to numerous discussions about the nature of this concept (Inbar-Lourie 2013). As this is a term not well-known in education (Elshawa et al. 2016), "several studies investigate what assessment literacy is, how it is conceptually defined, how it can be learned, and in what ways it can be productively developed as a stand-alone knowledge base" (Lam 2015, 170). Often embracing a range of skills and knowledge necessary for stakeholders to "deal with the new world of assessment into which we have been thrust" (Fulcher 2012, 115), assessment literacy often refers to teachers' understanding of what assessment methods to use and when to use them, to gather data about student achievement. Assessment literate teachers know what, why and how to assess and are aware of assessment problems, and how to prevent them (Stiggins 1991, 1995, 1999). They can recognize different purposes of assessment and can use them accordingly (Volante and Fazio 2007).

Language assessment literacy, specifically, appears to be a multi-layered entity (Inbar-Lourie 2013). Because of the broader lack of consensus within the professional testing community as to what constitutes assessment knowledge, defining it presents a major challenge. Despite considerable progress in the respective literature towards conceptualizing language assessment literacy, Hudaya (2017) argues that there is still no consensus about the necessary competencies that a language assessment literate person needs to have. According to Malone (2013), language assessment literacy refers to teachers' "familiarity with testing definitions and the application of this knowledge to classroom practices and to issues related to assessing language" (329). To be assessment literate, language teachers need to have assessment literacy skills, combined with language specific competencies (Jeong 2013). Pill and Harding (2013) further note that language assessment literacy indicates a repertoire of competences that enable individuals to understand, evaluate and, in some cases, create language tests and analyze test data. In this regard, ingredients of language assessment literacy can range from skills, such as item writing to complex issues, such as the relationship between language teaching approaches and language assessment (Inbar-Lourie 2013).

Language assessment literacy is often defined in the respective literature, in terms of three domains: knowledge, skills and principles (Giraldo 2018; Scarino 2013). In fact, Lam (2015) refers to "the acquisition of knowledge, skills and principles of test construction, test interpretation and use, test evaluation and classroom-based assessments, alongside the development of a critical stance about the functions of assessment within the wider educational context" (O'Loughlin 2013, 363). Knowledge reflects awareness of applied linguistics, theory and concepts, own language and the assessment context; skills relate to instruction, design for language assessment, educational measurement and technology; and principles refer to awareness of critical issues in language assessment (Giraldo 2018). Davies (2008) further suggests that knowledge relates to information about measurement and language, skills relate to the how-to or basic testing expertise, and principles relate to fairness, ethics and social consequences of language assessment. Teachers need to understand basic principles of sound assessment practice and distinguish this from unsound assessment (Paterno 2001; Bayat and Rezaei 2014), since knowledge of principles can make them critical towards assessment practice (Fulcher 2012; Scarino 2013). Hence, knowledge relates to proficiency and contextual considerations (Hudaya, 2017), skills relate to methodology and tools, such as item writing, statistics and analysis, while principles refer to assessment use, impact, fairness and ethics. Drawing on the domains of assessment literacy for teachers, Kahl et al. (2013) refer to them, as complementary to each other. However, other researchers place an emphasis on a certain domain. For instance, Vogt and Tsagari (2014) suggest that teachers mostly need skills to critique external tests.

At the same time, Giraldo (2018) suggests that two strands are involved in the assessment literacy notion: instruction and evaluation of valid assessment, and the appropriate use of tests and test results, along with the communication of the results to stakeholders. An assessment literate person can discern between excellent and poor-quality assessment and apply that knowledge to make informed inferences about student achievement (Stabler-Havener n.d.). Therefore, assessment literacy goes beyond knowledge about how to assess what students know and can do, towards interpreting assessment results and applying them to improve learning (Webb 2002). Teachers not only need to know what they are assessing and how they need to assess it according to specific purposes, but they also need to know what decisions they need to make to maximize learning (Djoub 2007). Beyond having the skills to define learning objectives to be assessed and make use of different assessment methods, assessment literate teachers can make inferences from analyzing students' work, provide constructive feedback to students and communicate the results of assessment (Sadler 1998). In fact, assessment literacy helps teachers to perceive, analyze, and use data on student performance to improve teaching (Bayat and Rezaei 2014). In this regard, teachers' assessment knowledge and competence can significantly impact teaching and learning (Elshawa et al. 2016).

3. Methodology

As already pointed out above, the PfL is a national longitudinal programme that aims to identify students at risk of being functionally illiterate by the end of compulsory education. The programme has a formative orientation–towards "assessment for learning"–and focuses on basic skills in two key competencies: language (Greek) and mathematics. These two competencies are considered to be the foundation for the development of other higher-order skills and also the prerequisites for accessing the broader curriculum (UNESCO 2006). This chapter focuses on language tests that assess three key skills: reading comprehension, writing and grammar. The PfL is also considered to be a monitoring mechanism towards equity in education by identifying struggling students early, so the necessary action is taken to prevent dropouts and ensure that all students acquire a minimum standard of education, before leaving school. Hence, to guide teaching and support decision-making at the classroom, school and education system levels for students that do not make adequate progress and face the risk of school failure, the programme provides critical information to stakeholders: policy makers, school leaders, teachers and parents. As the programme does not aim to evaluate teachers and schools, the negative washback effects on teaching and learning are limited.

4. The Assessment

4.1 Test Construction

Tests are developed every year by experts with teaching experience in primary education in the Greek language. The procedures employed are in line with the relevant literature (e.g. Downing and Haladyna 2006) and international studies such as PISA, and PIRLS. Item construction is based on specification tables, informed by the National Curriculum, as well as the literature on basic skills; these tables are valuable for evaluating content-related validity. Unlike common practice, specification tables are made publicly available. Therefore, although not involved in test construction, teachers are informed about the content of the tests.

Item writing follows rigorous iterative development and review processes. A range of item types such as multiple-choice, gap-filling, true/false, short-answer constructed response items and open-ended writing tasks is constructed. Sample test items are also made public, to allow students, teachers and parents to familiarize themselves with the items. Since the tests aim to identify students at the low end of the achievement distribution with precision, the difficulty distribution of the items does not match the whole distribution of student ability, while most items range from moderate to low difficulty.

Four parallel test versions for each year group are developed annually. For equating purposes, the tests for the same year group include common items, thus creating an overlap in test content. Common items are drawn from an item bank. So, new tests share common items with previous ones (internal anchor test). Equating is considered important, to establish comparability between alternate forms of tests built to the same content and statistical specifications by placing scores on a common scale, thus allowing the interchangeable use of scores on these forms (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 1999; Kolen and Brennan 1995). After equating, it does not make any difference which test form students have taken or when they have taken the test (Lord 1980). To equate tests, the fixed common item parameter method (FCIP) is used, with new items (and persons) placed on the established base scale (Kang and Petersen 2009).

4.2 Test Administration

The PfL language tests are administered to the full cohort of Year 3 (aged 8) and Year 6 (aged 11) students in all public primary schools (i.e., approximately 8000-9000 students in each year group). Teachers are responsible for indicating student exceptions from testing, on the grounds of limited proficiency in the Greek language and/or other difficulties e.g. special educational needs. Exceptions usually do not exceed 8% of the entire population.

In addition, teachers across all schools administer the language tests on fixed “national literacy days”, according to guidelines provided by the Centre for Educational Research and Evaluation (CERE). These guidelines aim to ensure both standardization of test administration, as well as valid and reliable results (Walklin 1990). According to Gronlund and Linn

(1990), tests can be successfully administered by any conscientious teacher, as long as the prescribed testing procedures are rigorously followed. Test administration is considered a key assessment competency skill for teachers. In this regard, teachers are entrusted with a very crucial role: to implement appropriate and fair testing conditions and ensure valid test results.

Finally, teachers are asked to complete a short questionnaire about the administration process, as well as test content. Specifically, teachers provide feedback on the adequacy of time provided for test completion, the level of difficulty of the tests for their students, their perceptions of the face validity of the tests, as well as suggestions for improvements. Teachers' perspectives on the assessment process improve the validity of the assessment (Ryan 2002). After administration, teachers are also responsible for returning the test materials to the CERE.

4.3 Test Coding and Data Analyses

To enhance reliability, marking is undertaken by external coders under standardized conditions (Kane 1992). Coders are usually new teachers, not yet employed to work in schools. Involving teachers in correcting and marking central assessments, may contribute substantially to their understanding of expected scoring criteria (Kane 2013, 198).

At this point, where performances on tests are translated into scores, scoring validity considerations arise. Detailed scoring rubrics are developed by the CERE as these allow for a high degree of consistency in scoring. To ensure uniformity of marking and to increase the reliability of scores, a training seminar on the use of rubrics and exemplars is conducted. Additionally, an online platform is provided to coders throughout the marking process for communication to enhance their capacity to apply scoring rubrics and marking criteria.

An important validity concern linked to scoring is related to interrater reliability: scores should not be a function of who scored the test, but rather a reflection of a student's performance in relation to the construct measured, as expressed through the scoring criteria (Slomp, Corrigan and Sugimoto 2014). To examine levels of agreement between coders, a double marking procedure is in place for 10% of the tests for each grade. Double-marking since 2013-2014 has indicated high interrater agreement.

The analysis of test data is undertaken by the CERE. Two well-known statistical models are employed for the analysis: the dichotomous Rasch model for constructing a common measurement scale for each construct and the Saltus model for setting a cut-off score for students "at risk".

The dichotomous Rasch model is employed for the examination of the quality of test data, the construction of item banks and the equation of tests from different test administrations. This simpler model from the Rasch family uses just one ability parameter for each person and one difficulty parameter for each item (based on persons' responses to items) to determine the probability of a certain person succeeding on a certain item (Bond and Fox 1999). The model builds a unidimensional scale (at interval level), along which persons and items are located according to their ability and difficulty measures; both person ability and item difficulty parameters have a common measurement unit (i.e., the logit scale). Two such scales are constructed for PfL language tests, one for each grade.

Before constructing the scales, rigorous statistical procedures are employed to assess the utility of the test data for Rasch analysis as the measurement properties (such as invariance, linear measures, estimates of precision, and devices for detecting misfit) of these models apply in practice, only when the model expectations or assumptions (unidimensionality, local statistical independence, etc.) are sufficiently met.

Validation analyses are also undertaken as the results of tests have an impact on many stakeholders (e.g., students, teachers, school leaders, and policy makers). Messick (1989) proposes validity as a unitary concept which is now widely accepted in language testing (Van der Walt and Steyn 2008). More importantly, Messick (1989) shifts validity from a property of the test to that of score interpretation and use, also introducing the notion of social consequences and ethics. Consequential validity refers to the adequacy and appropriateness of interpretations and actions based on test scores, supported by both empirical evidence and theoretical rationales (Messick 1989). This is relevant to the third component of language assessment literacy proposed by Davies (2008), namely principles (e.g., assessment use, impact, ethics and fairness). Two major threats to construct validity and score interpretation are construct underrepresentation (a test does not include all the essential elements of the construct in question) and construct-irrelevant variance (variance among scores attributed to factors that are irrelevant to the intended construct being measured). In the context of PfL, the item-person map and misfit statistics provided by the Rasch model are used to investigate and minimize sources of invalidity related to construct underrepresentation and construct-irrelevant variance as these can lead to erroneous results with negative consequences for the students.

Finally, the Saltus model is employed to identify two latent subgroups of persons, i.e., students “at risk” and students “not at risk” (Wilson 1989, Draney and Wilson 2007).

5. The Use of Assessment

5.1 Communication of Results to Schools

The final step involves communicating assessment results back to schools. Schools are not informed about students’ test scores, but rather about the students “at risk” (based on the Saltus model) or within the “grey zone”. The “grey zone” signifies uncertainty regarding students’ status based on test results which is related to two possibilities: their language abilities (as estimated by the Rasch model) are very close to the cut-off point (their estimated abilities falling within the confidence bands); and/or a high infit mean square value (>1.3) i.e., “misfits”. Various researchers raise concerns about high fit values as these might be a sign of “mismeasurement” or “invalidity” (Smith 1986). Individuals identified as “misfitting” may have been “mismeasured”, so their scores could be regarded as failing to provide valid measures of performance. For these students, the final decision remains at the school’s discretion. Teachers and school leaders are asked to make a decision as to whether students within the grey zone are “at risk”. This calls for the collection of evidence from different sources to make informed decisions.

After such decisions are made, the Departments of Primary and Secondary Education at the Ministry of Education are informed about the students who need support, while schools are responsible for informing parents and implementing intervention/support programmes for this population. Hence, policy makers and teachers need to address the individual needs of these students and provide effective education to help them reach a basic minimum level of skills.

5.2 Use of Results by Policy Makers

Policy makers need to have access to accurate, valid and reliable assessment data to implement policies to support students at risk (such as reallocating resources to students and schools with the greatest needs) and ensure a fairer and inclusive education system that makes the advantages of education available to all (OECD 2012).

The PfL provides policy makers with reliable and consistent information to monitor the worrying phenomenon of low basic skills.

Students "at risk" are quite likely to drop out of education upon completion of compulsory education, enter the workforce with low skills and unprepared, and face the risk of unemployment and/or low earnings (OECD 2012). Overall, the proportion of students at risk at Year 3 in the past four years (2014-2018) has ranged from 11% to 13% for the Greek language, while the respective percentages for Year 6 during the past four years (2004-2018) have ranged from 7% to 12% (Petridou and Neofytou 2015a, 2015b, 2016a, 2016b, 2017a, 2017b, 2018a, 2018b).

The CERE also provides policy makers with comparative data among the two cohorts (i.e., how many students who are identified at risk at Year 3 remain at risk at Year 6) to evaluate the effectiveness of support programmes. Comparative analyses of test data among the two cohorts also allow for the evaluation of the predictive validity of the tests.

5.3 Use of Results by Teachers

At the same time, in the context of PfL, teachers are provided with valid and reliable information regarding students who lack the basic skills to adapt and improve their teaching to meet learning needs. Assessment data are provided to schools to allow populations "at risk" to receive support.

A common criticism in relation to test data gathered in large-scale assessments is that feedback is not detailed enough to diagnose individual student needs (OECD 2013). To have the greatest impact, information is necessary not only on how the learner has done, but also on the specific steps needed to progress further. Therefore, the CERE provides some schools on a pilot basis with analytic reports on test performance for students at risk, indicating areas of weaknesses.

With regard to Year 3 students, school leaders are expected to develop school plans with, not only remedial, but also prevention strategies across all grades; classroom teachers also develop individualized student action plans (Ministry of Education and Culture 2013). Year 6 students receive extra support "in accordance with specially designed programmes in lower secondary education" (European Commission 2009, 51). Schools are asked to follow guidelines by the Department of Primary Education with useful teaching strategies and practices for teachers to effectively support these students. Since schools are expected to devise school-based literacy support programmes, as well as plans at the individual level (tailored, according to students' learning needs) that are regularly updated, monitored and evaluated, teachers need to be able to interpret and appropriately use the reports to make informed decisions.

6. National Literacy Assessment and Language Teaching

6.1 Enhancing language assessment literacy

Inbar-Lourie (2017) refers to the increase in the number of stakeholders from different disciplines, involved in language assessment and decision-making. In the psychometric era a gulf is developed between test developers and instructors, with the first responsible for large-scale testing and the second responsible for the instruction of learners (Malone 2008). According to Taylor (2013), researchers and test makers are at the core of this process, language teachers are at the intermediary level and policy makers and the general public at the peripheral level.

In this regard, stakeholder groups have different profiles in terms of assessment literacy (Taylor 2013). Taylor (2009) points out that language assessment literacy extends beyond the teaching profession and suggests differential profiles for four groups: test writers, classroom teachers, university administrators and language testers. Although the language assessment literacy construct embraces all dimensions, such as knowledge of theory, technical skills, principles and concepts, language pedagogy, sociocultural values, local practice, personal beliefs and scores/decision-making, different stakeholders need to address each one of these to a different degree (Taylor 2013). Giraldo (2018) refers to "different shades of meaning for various people directly or indirectly involved in language assessment" (184). Inbar-Lourie (2017) also emphasizes that future directions point at a move towards situated differentials rather than unified conceptualization in the form of language assessment literacies.

This chapter places special interest on teachers, and what the PfL programme involves for them, with regard to the development of language assessment literacy. To begin with, in terms of the stages and processes outlined above, the PfL appears to resemble a hybrid model of national testing; test construction and marking remain the responsibility of the educational authorities while test administration is undertaken by teachers. Moreover, teachers are expected to review the results, communicate them to parents and use them in teaching. Therefore, within the context of this programme, language teachers are mostly considered as users of test data for student learning. Assessment data are made available to these users, under the assumption that these become useful for instruction or other purposes, such as communication and evaluation, e.g. informing parents about students' activity or student appraisal. As Malone (2013) argues, properly implemented assessment provides teachers with important information about student performance and about the extent to which learning objectives have been attained. However, to use test data for student learning, teachers need to translate results into actions that support learning (Hopster-den-Otter et al. 2017). In fact, as Razavipour (2013) discusses, the validity of tests depends on the interpretations attempted, the inferences drawn and decisions as to the uses of the test scores. Apparently, the PfL screens individual students who are not performing at the expected level and may need additional or individualized assistance. In this regard, as an early literacy assessment, the programme provides service to teachers as they attempt to prevent reading difficulty or failure (Helman 2005). Teachers are expected to use the results, along with other language assessment data they collected themselves, to inform their language teaching approaches and design effective instructional and enriched literacy-learning experiences, involving appropriate reading and writing activities for the students "at risk".

At the same time, beyond the traditional paradigm "researchers produce knowledge and policy-makers consume this" and the linear notion of "knowledge transfer from theorists-researchers to practitioners-policy makers" (Carasco et al. 2001, 258), the PfL can enhance synergies between testers and teachers (Brindley 2001, Brown and Bailey 2008, Davies 2008). It is often the case that the interests of language testing experts may not be of interest to language instructors (Malone 2013, 343). Since this national central assessment programme allows different stakeholders involved in language testing and instruction to work together, the PfL moves towards a direction where teachers, testing experts and policy makers try to achieve common goals. Bringing down the divide between testing experts and teachers can, not only nurture teacher language assessment literacy but also promote an assessment culture which values different sources of data as equally significant for the advancement of learning goals (Inbar-Lourie 2008). Although roles and responsibilities might differ, negotiations of meanings between researchers, policy makers and teachers can contribute to assessment literacy development for all stakeholders.

6.2 Acting further

Apparently, it is necessary to develop further leadership capacity for teachers and leaders to examine data, critically discuss them and develop action plans based on such data; "what is counted, what is omitted, who is involved and the categories and procedures developed and enacted all matter when seeking to understand the nature and effects of using numbers and statistics to make sense of the world" (Hardy 2013, 68). As Earl and Fullan (2003) note, sometimes leaders are experiencing "tensions and dilemmas as a result of data about their schools, especially when such data are collected centrally as part of a large-scale reform agenda" (392). Teachers need to be aware of the claims that can be made based on the PfL test results but also to have the capacity to ask the right questions. In this regard, the professional preparation of language teachers could involve a sharper focus on testing paradigms and establish that teachers have fundamental competences in assessment (Yin 2010).

At the same time, as Giraldo (2018) argues, knowledge, skills and principles in language assessment coexist with teachers' ways of thinking and acting upon the assessment. According to Hardy (2013), "how national testing actually plays out in relation to literacy practices in school settings is not necessarily easily or readily captured" (76). Scarino (2013) further refers to the importance of teachers' interpretative frameworks, while Borg (2003) discusses teacher cognition, i.e. what teachers think, know, and believe and the relationships of these mental constructs to what teachers do in the language teaching classroom. In the case of the PfL, despite guidelines to schools about support for students "at risk", the extent to which these protocols are effective in raising literacy in Cyprus still remains unknown. Hence, light could be shed onto the actual school and teacher practices to support students "at risk", the problems involved, the professional development needs, as well as the usefulness and effectiveness of the strategies employed. It would certainly be interesting to extend research into teachers' and leaders' meanings of the particular programme. Large-scale standardized testing is sometimes seen as "professionally debilitating" because it devalues teachers' pedagogical skills, and reflects state authority and models of control over a body of content and pedagogy that should belong to the teachers; also, testing removes the judgemental right of appraisal in education (Gambell and Hunter 2004). Research studies involving a focus on teachers and their knowledge and attitudes to assessment practices are widely reported in various national settings e.g. Baker and Riches (2017) discuss teachers' language literacy development within the Haitian context, Berry et al. (2019) focus on teachers in the United Kingdom, France and Spain, while Yan et al. (2018) present work on teachers in China. Therefore, teachers' understanding of standardized national literacy testing could be further investigated, as well as professional reflections on the relevance and impact of this testing on classroom teaching and learning (Dreher 2012). Analysing teachers' and leaders' responses to the PfL would help the Cyprus educational system to obtain a qualitative account of what happens in schools and why.

7. Concluding Remark

This chapter has argued about the potential of this particular national assessment programme to enhance teachers' language assessment literacy and allow testing experts and teachers to work together towards a common goal, i.e., the identification and support of students "at risk" in the Cypriot context. At the same time, the chapter has highlighted possible directions in which the PfL can further develop to impact language teaching more effectively.

In contrast to external assessments linked to accountability, national assessment programmes with a formative orientation such as the PfL can contribute to the development of collaborative interactions among stakeholders, also empowering teachers to develop language assessment literacy (Grabowski and Dakin 2014). That said, the possible impact of such national programmes needs to be explored, against the broader contextual educational framework in which they are developed. As the OECD (2013) suggests:

> Overall, while there is some evidence that data from large-scale assessments are being used successfully to identify students' strengths and weaknesses, to change regular classroom practice or to make decisions about resource allocation they need to be embedded in broader, more comprehensive assessment systems that include a range of summative and formative assessments, curriculum tasks, instructional tools, and professional development that helps teachers understand which assessment information is most appropriate for a particular purpose. (159)

8. References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 1999. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Bayat, Khadijeh, and Amir Rezaei. 2014. "Importance of Teachers' Assessment Literacy." *International Journal of English Language Education* 3(1):139-146.

Baker, Beverly, and Caroline Riches. 2017. "The Development of EFL Examinations in Haiti: Collaboration and Language Assessment Literacy Development." *Language Testing* 35(4):557 –581.

Berry, Vivien, Susan Sheehan, and Sonia Munro. 2019. "What Does Language Assessment Literacy Mean to Teachers?" *ELT Journal* 73(2):113-123.

Bond, Trevor, and Christine Fox. 1999. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences.* Mahwah, NJ: Erlbaum.

Borg, Simon. 2003. "Teacher Cognition in Language Teaching: A Review of Research on what Language Teachers Think, Know, Believe, and Do." *Language Teaching* 36(2):81-109.

Brindley, Geoff. 2001. "Language Assessment and Professional Development." In *Experimenting with Uncertainty: Essays in Honour of Alan Davies*, edited by Catherine Elder, Alison Brown, Kathryn Hill, Noriko Iwashita, Tom Lumley, Tim McNamara, and Kieran O'Loughlin, 126-136. Cambridge: Cambridge University Press.

Brown, James Dean, and Kathleen Bailey. 2008."Language Testing Courses: What are they in 2007?" *Language Testing* 25(3):349-383.

Carasco, Joseph, Nancy Clair, and Lawrence Kanyike. 2001. "Enhancing Dialogue among Researchers, Policy Makers, and Community Members in Uganda: Complexities, Possibilities, and Persistent Questions." *Comparative Education Review* 45(2):257-279.

Catts, Hugh, Marc Fey, Xuyang Zhang, and Bruce Tomblin. 2001. "Estimating the Risk of Future Reading Difficulties in Kindergarten Children: A Research-based Model and its Clinical Implementation." *Language, Speech, and Hearing Services in the Schools* 32(1):38-50.

Coombe, Christine, Mashael Al-Hamly, and Salah Troudi. 2009. "Foreign and Second Language Teacher Assessment Literacy: Issues, Challenges and Recommendations." *Research Notes* 38:14-18.

Davies, Alan. 2008. "Textbook Trends in Teaching Language Testing." *Language Testing* 25(3):327-347.

Davies, Michael, and Stephen Elliott. 2012. "Inclusive Assessment and Accountability: Policy to Evidence-based Practices." *International Journal of Disability, Development and Education* 59(1):1-6. doi: 10.1080/1034912X.2012.654932.

Demetriou, Demetris, and Leonidas Kyriakides. 2006. "Investigating the Generalisability of Models of Educational Effectiveness: A Secondary Analysis of PISA Study." Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 7–11.

Djoub, Zineb. 2007. "Assessment Literacy: Beyond Teacher Practice." In *Revisiting EFL Assessment: Critical Perspectives (Series Second Language Learning and Teaching),* edited by Rahma Al-Mahrooqi, Christine Coombe, Faisal Al-Maamari and Vijay Thakur, 9-27. New York: Springer.

Downing, Steven, and Thomas Haladyna. 2006. *Handbook of Test Development*. Lawrence: Erlbaum Associates.

Draney, Karen, and Mark Wilson. 2007. "Application of the Saltus Model to Stage-like Data: Some Applications and Current Developments." In *Multivariate and Mixture Distribution Rasch Models,* edited by Mathias von Davier, and Claus H. Carstense, 119-130. New York: Springer.

Dreher, Karen. 2012. "Tests, Testing Times and Literacy Teaching." *Australian Journal of Language and Literacy* 35(3):334-352.

Earl, Lorna, and Michael Fullan. 2003. "Using Data in Leadership for Learning." *Cambridge Journal of Education* 33(3):383-94.

Elshawa, Niveen, Chan Swee Heng, Ain Nadzimah Abdullah, and Sabariah Rashid. 2016. "Teachers' Assessment Literacy and Washback Effect of Assessment." *International Journal of Applied Linguistics and English Literature* 5(4):135-141.

Eurydice. 2016. "Cyprus: Overview." Accessed February 13, 2019. https://webgate.ec.europa.eu/fpfis/mwikis/eurydice/index.php/Cyprus:Redirect.

European Commission. 2009. "National Testing of Pupils in Europe: Objectives, Organisation and Use of Results." *Brussels: Education, Audiovisual and Culture Executive Agency* (EACEA P9 Eurydice). Accessed February 13, 2019. http://eacea.ec.europa.eu/education/eurydice/documents/thematic_reports/109en.pdf.

Faubert, Brenton. 2012. "A Literature Review of School Practices to Overcome School Failure." *OECD Education Working Papers*: 68.

Fulcher, Glenn. 2012. "Assessment Literacy for the Language Classroom." *Language Assessment Quarterly* 9(2):113-132.

Gambell, Trevor, and Darry Hunter. 2004. "Testing Scoring of Large-scale Assessment: Professional Development or Debilitation?" *Journal of Curriculum Studies* 3(6):697-724.

Giraldo, Frank. 2018. "Language Assessment Literacy: Implications for Language Teachers." *Profile: Issues in Teachers' Professional Development* 20(1): 179-195. https://doi.org/10.15446/profile.v20n1.62089.

Grabowski, Kirby, and Jee Wha Dakin. 2014. "Test Development Literacy." In *The Companion to Language Assessment*, edited by Antony John Kunnan, 751-768. Chichester, West Sussex: John Wiley & Sons, Inc.

Gronlund, Norman, and Robert Linn. 1990. *Measurement and Evaluation in Teaching* (6th ed.). New York: Macmillan Publishing Company.

Hardy, Ian. 2013. “Testing that Counts: Contesting National Literacy Assessment Policy in Complex Schooling Settings.” *Australian Journal of Language and Literacy* 36(2):67-77.

—. 2015. “A Logic of Enumeration: The Nature and Effects of National Literacy and Numeracy Testing in Australia.” *Journal of Education Policy* 30(3):335-362.

Helman, Lori. 2005. “Using Literacy Assessment Results to Improve Teaching for English-language Learners.” *The Reading Teacher* 58(7):668-677.

Hill, Kathryn. 2017. “Language Teacher Assessment Literacy–Scoping the Territory.” *Papers in Language Testing and Assessment* 6(1): iv-vii.

Hipwell, Pat, and Valentina Klenowski. 2011. “A Case for Addressing the Literacy Demands of Student Assessment.” *Australian Journal of Language and Literacy* 34(2):127-146.

Hopster-den-Otter, Dorien, Saskia Wools, Theo T. J. H. M. Eggen and Bernard Veldkamp. 2017. “Formative Use of Test Results: A User’s Perspective.” *Studies in Educational Evaluation* 52:12-23.

Hudaya, Diah Wulansari. 2017. “Teachers’ Assessment Literacy in Applying Principles of Language Assessment.” In *Proceedings of the 1st Education and Language International Conference, Center for International Language Development of Unissula*, 247-260. http://jurnal.unissula.ac.id/index.php/ELIC/article/viewFile/1238/946.

Inbar-Lourie, Ofra. 2008. “Constructing a Language Assessment Knowledge Base: A Focus on Language Assessment Courses.” *Language Testing* 25(3):385-402. https://doi. org/10.1177/0265532208090158.

—. 2013. “Guest Editorial to the Special Issue of Language Assessment Literacy.” *Language Testing* 30(3):301-307.

—. 2017. “Language Assessment Literacy.” In *Encyclopedia of Language and Education: Language Testing and Assessment*, vol. 7, edited by Elana Shohamy, and Nancy Hornberger, 257-270. New York: Springer. DOI 10.1007/978-3-319-02261-1_19.

Jeong, Heejeong. 2013. “Defining Assessment Literacy: Is It Different for Language Testers and Non-language Testers?” *Language Teaching* 30(3):345-362.

Jin, Yan. 2010. “The Place of Language Testing and Assessment in the Professional Preparation of Foreign Language Teachers in China.” *Language Testing* 27(4):555-584.

Karagiorgi, Yiasemina, and Alexandra Petridou. 2019.“National Literacy Assessment: The Identification of Students ‘At Risk’ in Cyprus.” *Research Papers in Education* 34(1):1-13.

https://doi.org/10.1080/02671522.2017.1329341.

Kahl, Stuart, Peter Hofman, and Sara Bryant. 2013. *Assessment Literacy Standards and Performance Measures for Teacher Candidates and Practising Teachers*. Prepared for the Council for the Accreditation and Educator Preparation (CAEP). Accessed February 28, 2019. http://caepnet.org/~/media/Files/caep/standards/assessment-literacy-in-teacher-preparati.pdf.

Kane, Michael. 1992. "An Argument-based Approach to Validity." *Psychological Bulletin* 112:527-535.

—. 2013. "Validating the Interpretations and Uses of Test Scores." *Journal of Educational Measurement* 50(1):1-73.

Kang, Taehoon, and Nancy Petersen. 2009. Linking Item Parameters to a Base Scale (No. 2009-2): ACT.

Kolen, Michael, and Robert Brennan. 1995. *Test Equating: Methods and Practices*. New York: Springer.

Lam, Ricky. 2015. "Language Assessment Training in Hong Kong: Implications for Language Assessment Literacy." *Language Testing* 32(2):169-197.

Lord, Frederic. 1980. *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum.

Malone, Margaret. 2008. "Training in Language Assessment." In *Encyclopedia of Language and Education: Language Testing and Assessment*, vol. 7, edited by Elana Shohamy, and Nancy Hornberger, 225-239. New York: Springer. https://doi.org/10.1007/978-0-387-30424-3_178.

Malone, Margaret. 2013. "The Essentials of Assessment Literacy: Contrasts Between Testers and Users." *Language Testing* 30(3):329-344.

Messick, Samuel. 1989. "Validity." In *Educational Measurement* edited by R. L. Linn, 13-104. New York: Macmillan.

Ministry of Education and Culture. 2013. "Supporting Students at Risk for Functional Illiteracy." Accessed 10 June 2016. http://egkyklioi.moec.gov.cy/Data/dde3723a.pdf.

—. 2014. "Annual Report." Accessed 10 June 2016. http://www.moec.gov.cy/en/annual_reports/annual_report_2014_en.pdf.

O'Loughlin, Kieran. 2013. "Developing the Assessment Literacy of University Proficiency Test Users." *Language Testing* 30(3):363-380.

OECD. 2012. "Equity and Quality in Education: Supporting Disadvantaged Students and Schools." OECD Publishing. Accessed 30 May 2016. http://dx.doi.org/10.1787/9789264130852-en.

—. 2013. "Student Assessment: Putting the Learner at the Centre". In *Synergies for Better Learning: An International Perspective on Evaluation and Assessment.* OECD Publishing. https://doi.org/10.1787/9789264190658-7-en.

Paterno, J. 2001. "Measuring Success: A Glossary of Assessment Terms." In *Building Cathedrals: Compassion for the 21st Century*.

Petridou, Alexandra, and Elsa Neofytou. 2015a. *The 2014-2015 Report on the Dimensions of Functional Illiteracy for Pupils aged Eight in Cyprus* [Research Report]. Nicosia: Cyprus Centre for Educational Research and Evaluation.

—. 2015b. *The 2014-2015 Report on the Dimensions of Functional Illiteracy for Pupils aged Eleven in Cyprus* [Research Report]. Nicosia: Cyprus Centre for Educational Research and Evaluation.

—. 2016a. *The 2015-2016 Report on the Dimensions of Functional Illiteracy for Pupils aged Eight in Cyprus* [Research Report]. Nicosia: Cyprus Centre for Educational Research and Evaluation.

—. 2016b. *The 2015-2016 Report on the Dimensions of Functional Illiteracy for pupils aged Eleven in Cyprus* [Research Report]. Nicosia: Cyprus Centre for Educational Research and Evaluation.

—. 2017a. *The 2016-2017 Report on the Dimensions of Functional Illiteracy for Pupils aged Eight in Cyprus* [Research Report]. Nicosia: Cyprus Centre for Educational Research and Evaluation.

—. 2017b. *The 2016-2017 Report on the Dimensions of Functional Illiteracy for Pupils aged Eleven in Cyprus* [Research Report]. Nicosia: Cyprus Centre for Educational Research and Evaluation.

—. 2018a. *The 2017-2018 Report on the Dimensions of Functional Illiteracy for Pupils aged Eight in Cyprus* [Research Report]. Nicosia: Cyprus Centre for Educational Research and Evaluation.

—. 2018b. *The 2017-2018 Report on the Dimensions of Functional Illiteracy for Pupils aged Eleven in Cyprus* [Research Report]. Nicosia: Cyprus Centre for Educational Research and Evaluation.

Pill, John, and Luke Harding. 2013. "Defining the Language Assessment Literacy Gap: Evidence from a Parliamentary Inquiry." *Language Testing* 30(3): 381-402.

Popham, James. 2009. "Assessment Literacy for Teachers: Faddish or Fundamental?" *Theory into Practice* 48(1):4-11.

Razavipour, Kioumars. 2013. "Assessing Assessment Literacy: Insights from a High-stakes Test." *Research in Applied Linguistics* 4(1):111-131.

Ryan, Katherine. 2002. "Assessment Validation in the Context of High-stakes Assessment." *Educational Measurement: Issues and Practice* 22: 7-15.

Sadler, Royce D. 1998. "Formative Assessment: Revisiting the Territory." *Assessment in Education*: *Principles, Policy and Practice* 5(1):77-84.

Scarino, Angela. 2013. "Language Assessment Literacy as Self Awareness. Understanding the Role of Interpretation in Assessment and in Teacher Learning." *Language Testing* 30(3):309-327.

Slomp, David, Julie Corrigan, and Tamiko Sugimoto. 2014. "A Framework for Using Consequential Validity Evidence in Evaluating Large-scale Writing Assessments." *Research in the Teaching of English* 48(3):276-302.

Smith, Richard. 1986. "Person Fit in the Rasch Model." *Educational and Psychological Measurement* 46(2):359-372.

Stabler-Havener, Michelle. (n.d.). "Defining, Conceptualizing, Problematizing, and Assessing Language Teacher Assessment Literacy." *Working Papers in Applied Linguistics and TESOL* 18(1):1-22.

Stiggins, Richard. 1991. "Assessment Literacy." *Phi Delta Kappan* 72:534-539.

—. 1995. "Assessment Literacy for the 21st Century." *Phi Delta Kappan,* 77(3):238-245.

—. 1999. "Assessment, Student Confidence and School Success." *Phi Delta Kappan* 81(3):191-198.

Taylor, Lynda. 2009. "Developing Assessment Literacy." *Annual Review of Applied Linguistics* 29: 21-36.

—. 2013. "Communicating the Theory, Practice and Principles of Language Testing to Test Stakeholders: Some Reflections." *Language Testing* 30(3):403-412.

Thomas, Kelly, and Douglas Huffman. (n.d.). "Navigating the Challenges of Helping Teachers Use Data to Inform Educational Decisions," 94-102. Accessed March 19, 2019. https://files.eric.ed.gov/fulltext/EJ1055020.pdf.

UNESCO (2006) Literacy for Life-Education for All. UNESCO Publishing France. Retrieved from: https://unesdoc.unesco.org/ark:/48223/pf0000141639

Unsworth, Len, and Eveline Chan. 2009. "Bridging Multimodal Literacies and National Assessment Programs in Literacy." *Australian Journal of Language and Literacy* 32(3):245-257.

Van der Walt, Johann, and Faans Steyn. 2008."The Validation of Language Tests." *Stellenbosch Papers in Linguistics* 38:191-204.

Vogt, Karin, and Dina Tsagari. 2014. "Assessment Literacy of Foreign Language Teachers. Findings of a European Study." *Language Assessment Quarterly* 11(4):374-402.

Volante, Louis, and Xavier Fazio. 2007. "Exploring Teacher Candidates' Assessment Literacy: Implications for Teacher Education Reform and Professional Development." *Canadian Journal of Education* 30:749-770.

Walklin, Les. 1990. *Teaching and Learning in Further and Adult Education*. Cheltenham: Stanley Thornes.

Webb, Norman. 2002. "Assessment Literacy in a Standards-based Educational Setting." Paper at the American Educational Research Association. New Orleans, Louisiana.

Wilson, Mark. 1989. "Saltus: A Psychometric Model of Discontinuity in Cognitive Development." *Psychological Bulletin* 105(2):276-289.

Yan, Xun, Cong Zhang, and Jason Jingson Fan. 2018. "Assessment Knowledge is Important, but…': How Contextual and Experiential Factors Mediate Assessment Practice and Training Needs of Language Teachers." *System* 74:158-168.

Yin, Muchun. 2010. "Understanding classroom language assessment through teacher thinking research". *Language Assessment Quarterly* 7(2):175–194.

PART III:

EXPANDING LANGUAGE ASSESSMENT LITERACY

CHAPTER EIGHT

INSIDE THE BLACK BOX OF L2 READING ABILITY MEASUREMENT: INTERFACES BETWEEN ASSESSMENT LITERACY AND VALIDITY

ASMA MAAOUI AND DINA TSAGARI

Attention paid to teachers' conceptions of assessment and assessment literacy in second language assessment has gained momentum in the last decade. Underpinned by the consideration of a broad view of validity, it has now been recognized that Language Assessment Literacy (LAL) should be studied from a culturally-situated perspective. This study examined teacher- and context-related factors in reading assessment in Tunisian higher education within the interface of LAL and validity. Data were collected through semi-structured interviews with three item writers of a reading achievement test to explore their assessment literacy. The findings of the study revealed issues of test design as a crucial step in determining the validity of arguments about the inferences drawn in relation to second language reading. The results disclosed fundamental intrinsic and extrinsic factors that shaped the teachers' assessment literacy at the individual level. The study also led to the identification of challenges caused by inadequate teacher assessment literacy underlying the test design process. The findings have important implications for assessment literacy development and for stakeholders' responsibility in benchmarking best practices in teacher-led high-stakes assessment.

1. Introduction

The area of language assessment literacy (LAL) has received unprecedented attention over the last few decades (Hasselgreen 2008; Taylor 2009, 2013; Vogt & Tsagari 2014; Yan 2010). Concern about LAL emerged as an outcome of the major shift from a traditional assessment paradigm "comparing students with their peers based on achievement to an era when we compare student performance to pre-set standards" (Stiggins 2006, 3). This has resulted in growing needs for insights into the nature of LAL and the implications it may have on high-quality classroom assessment thus providing an impetus for the development of new pedagogically-oriented assessment materials (Inbar-Lourie 2013; Tsagari et al. 2017) and teacher training programs to enhance LAL (Fulcher 2012). The role assigned to teachers as key agents in classroom assessment places a heavy responsibility on them to acquire LAL as a fundamental professional component (Inbar-Lourie 2008; Looney, Cumming, van Der Kleij, & Harris 2018; Stiggins 2001; Xu & Brown 2016). LAL allows teachers to be conversant in the social context of assessment including the construction, use and evaluation of language tests. Unfortunately, assessment has remained a professional area of "marginal relevance to everyday classroom concerns" (Brindley 2001, 127) for decades. This should draw attention to the potential dangers of *uninformed* assessment practices or what Popham (2004) describes as "professional suicide".

Finally, despite developments in LAL research, studies outside the Western educational context (Sultana 2019) have remained rather scarce. To date, there is a dearth of assessment research in a country like Tunisia (Athimni 2018; Hidri 2016; Maaoui 2016). Research in this context has demonstrated that teachers have "wrong and conflicting assessment conceptions" (Hidri 2016, 33). Motivated by these issues, this study will attempt to delineate the assessment landscape in a Tunisian higher education context while underscoring LAL versus valid assessment demands for reading achievement purposes.

2. Conceptual Framework

With the wide recognition of the role of the teacher as assessor (Miller, Linn, & Gronlund 2009; Rea-Dickins 2004), research has sought to determine the nature and role of LAL in classroom assessment. An "assessment-literate" (Brown 2008, 286) or "assessment-capable teacher" (Hill, Cowie, Gilmore, & Smith 2010; Smith, Hill, Cowie, & Gilmore 2014; DeLuca & Johnson 2017) is described as "one who creates, chooses, administers, interprets, responds to, records, and reports assessment information" adequately and usefully (Brown 2008, 286). Studies of "teachers' professional judgement" (Allal 2013; Allal & Lopez 2014) within a socio-cognitive (Weir 2005) or a sociocultural perspective (Constantinou, Crisp, & Johnson 2017; Inbar-Lourie 2016; Scarino 2013) research paradigm have revealed the complexity of "assessment beliefs" and their common convergence with teaching practices and assessment policies (Looney et al. 2017) thus bringing "assessment identity" and its dynamic and experiential nature to light (Scarino 2013). Concurrent with the reconceptualization of assessment literacy (AL), Xu and Brown (2016) relied on a scoping review of 100 studies to develop the framework of Teacher Assessment Literacy in Practice (TALiP). This conceptual model comprised six interrelated components including a teacher knowledge base, conceptions of assessment and emotional interactions with assessment further accounting for the various dimensions of AL. Teacher beliefs are presented as "an interpretive and guiding framework by which they mediate their uptake of theoretical knowledge and its implementation" (Xu & Brown 2016, 156) that are shaped by overarching conceptions of teaching and learning (Brown 2008).

The nature of assessment literacy has witnessed major developments in educational and language assessment research. First coined by Stiggins (1991) as "the ability to understand, analyze and apply information on student performance to improve instruction" (Falsgraf 2005, 6), it has been established as a distinct entity. Debates about the concept were initiated with attention paid to teacher views, knowledge, skills (Fulmer, Lee, & Tan 2015; Popham 2004) and conceptions of assessment (Brown, Irving, & Keegan 2007). There is an overt consensus in the literature on the need for adequate *knowledge* of assessment principles (Popham 2004; Stiggins 1991, 2001), skills-based know-how (Taylor 2009), "competencies" (Pill & Harding 2013) and ability to put them into practice (DeLuca & Bellara 2013; DeLuca & Johnson 2017) for teachers in regular language classrooms. In this study, LAL will be operationalized as the teachers' (1) assessment *knowledge,* and (2) *skills* (Brindley 2001; Fulcher 2012; Harding & Kremmel 2016; Popham 2009; Vogt & Tsagari 2014, 377). This knowledge should be translated into an ability to design, develop and critically evaluate tests by aligning "assessment purposes, content and methods" (Xu & Brown 2016, 156).

3. LAL Research

Despite the turn in language assessment, LAL research is still in its infancy. Little is known of teacher-led assessment processes (Black & Wiliam 2010). The ostensibly present teacher assessment professionalism has also been put into question. Research in educational assessment at large has shown that only a minority of teachers are prepared "to face the challenges of classroom assessment because they have not been given the opportunity to learn to do so" (Stiggins 2004, 762). Assessment purposes were found to be often implicit, unstated or simply unknown (Maaoui 2016). Most importantly, early AL research provided empirical evidence (López Mendoza & Bernal Arandia 2009; Yan 2010; Fulcher 2012; Hasselegreen, Carlsen, & Helness 2004) mainly of a lack of adequate assessment knowledge and skills. Studies examining teacher-based assessment (Davison 2004; Davison & Leung 2009; Rea-Dickins 2004; 2007) have pointed out a range of language measurement problems including low levels of agreement between teachers regarding the construct assessed. Findings (e.g., Brindley 2001) have shown a disparity in the way teachers operationalize and interpret assessment descriptors. Davison's (2004) comparison of two groups of teachers in three sociocultural contexts showed individual teachers' diverging use of theoretical assessment criteria in their test design process. However, more LAL research is needed to pinpoint the nature of the concept itself (Pill & Harding 2013; Scarino 2017; Taylor 2009) and the various ramifications of insufficient LAL in teacher-based assessment (Vogt & Tsagari 2014; Tsagari 2016; Tsagari & Vogt 2017). This kind of research can be compared to a discovery of what is in "the black box" of classroom assessment (Shohamy 1998; Black & Wiliam 2010).

Methodologically, the first line of research sought mainly to confirm the psychometric properties of conceptions of assessment through correlational studies with pre-service and in-service teachers. Structural equation modelling also served to confirm an association between teachers' assessment conceptions and practices (Brown, Kennedy, Fok, Chan, & Yu 2009). Surveys (Fulcher 2012; Hasselgreen, Carlsen, & Helness 2004; Huhta, Hirvalä, & Banerjee 2005) aimed to determine assessment training needs in different educational contexts. However, "all of these studies suffer from their utilization of primarily closed-response items, which lend themselves only to quantitative treatment" (Fulcher 2012, 118). Researchers have predominantly relied on inventories of the purposes of assessment (Brown 2006, 2008), LAL inventories and questionnaires (Popham 2014), large-scale conceptions of assessment (Brown 2006) and LAL surveys (Vogt & Tsagari 2014). The structured nature of the evidence gathered from teachers probably prevented any inter-personal, context-specific differences among teachers' LAL. Researchers developed various LAL inventories and questionnaires (Brown 2006; Popham 2014) but the main problem associated with this line of research was "the construction of the a priori categories for analysis" (Fulcher 2012, 127). Despite the usefulness of this line of LAL research, this area needs more in-depth qualitative research.

To gain more insights into the nature of LAL, recent research has attempted to improve data elicitation methods. This new research orientation has examined LAL in regular classrooms from a more qualitative perspective. A dearth of research has relied on interviews with teachers to explore the nature of the construct and explore its relation to classroom-based assessment (Deneen & Brown 2016; Sellan 2017). Exceptionally, while accounting for such methodological limitations, Vogt and Tsagari (2014) and Tsagari and Vogt (2017) combined quantitative and qualitative methods. Even though Vogt and Tsagari's (2014) study replicated the questionnaire from Hasselgreen et al. (2004) to gather LAL data from seven European countries, it has taken a step forward by conducting follow-up interviews with a sample from the participants. The results from both instruments confirmed early conclusions about training as a determining factor in developing LAL. The study makes suggestions for the development of an in-service teacher training "format" involving cooperation with teachers that could cater for their needs (Vogt & Tsagari 2014, 392). In more recent research, Hill (2017) developed a framework to help teachers articulate descriptions of their assessment practices while Scarino (2017) and Sellan (2017) examined the development of teacher AL adopting an "interpretive research approach" about assessment from within. Tsagari and Vogt's (2017) mixed-methods study of primary and secondary teachers from three European countries delved into individual teachers' profiles, LAL perceptions and their training needs through semi-structured interviews. The findings revealed that "the informants' perceived LAL profiles are not sufficient for the assessment activities they have to accomplish in their professional field" (7). To cope with the demands of testing, the teachers tended to discover language testing and assessment through practice while relying on counsellors, other teachers and available assessment materials.

Overall, the research literature has shown that LAL studies have clearly attended to the various factors shaping teachers' assessment knowledge, views, or practices (Fulmer et al. 2015) and that LAL can be a determining factor in language test design. Therefore, this study aims to explore teachers' LAL, including the various characteristics of the specific educational context within which it is studied and the link between LAL, test design and validity. The following section will present the study in terms of purpose and design. This will be followed by a description of contextual factors interacting with the main construct investigated.

4. Theoretical Framework

Addressing LAL in achievement testing among other assessment practices, the present study combines three theoretical dimensions in a framework described in Figure 1. Relying on current validity discussions (Moss 2015), validity is viewed as a broad and complex construct predominantly dependent on teachers' LAL which, in itself, is a determining factor in test design and use. Within this *continuing* debate on the nature of validity (Bachman & Palmer 2010; Kane 2006; Moss 2015), this study undertakes to account for the interface between LAL and "*test purpose*, test design and validation" in a more explicit way (Fulcher 2012, 1). Taking Kane's (2006) argument-based validity a step further, an "assessment use argument" (Bachman and Palmer 2010) is particularly useful as it proposes that research should scrutinize validity as early as test design (see Fulcher 2017). Within this vein, recent assessment research (Moss 2015) has recently started to make calls for a re-conceptualization of validity to examine "actual interpretations and uses by professionals in educational contexts" of tests and related data (1). Recent validation research (Davison 2004; Kane 2013; Newton, 2016; Weir 2005) equally calls attention to the necessity of re-considering validity mainly in contexts of high-stakes teacher-led assessment. Therefore, the framework incorporates (1) LAL, (2) test design, and (3) validity as interrelated facets in teacher-based assessment.

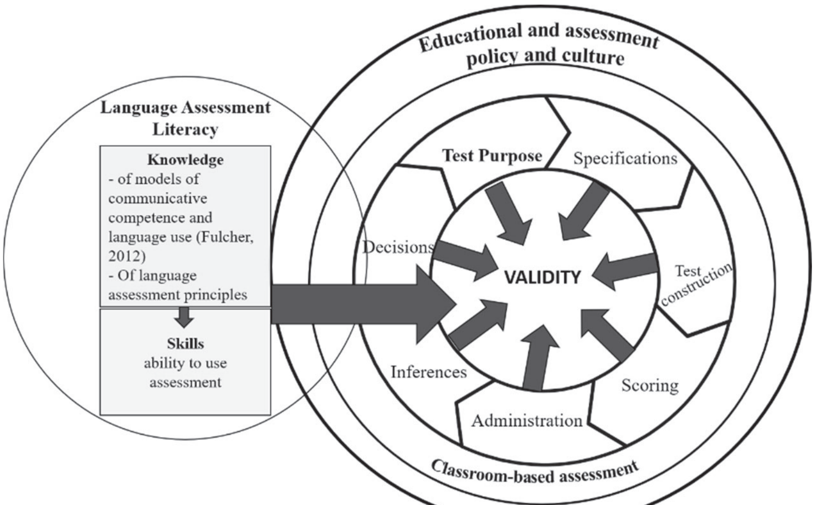


Figure 1. Interfaces between LAL and Validity
(based on Vogt & Tsagari 2014; Fulcher 2010, 2017; Moss 2015)

This is justified by the warning of any “design chaos” at the onset of the test development process as this may also lead to “validity chaos" at the end (Fulcher 2010; Fulcher & Davison 2009). In addition to the need for knowledge of theoretical models of language, teachers should be able to state an overt *test purpose* elucidating the “domains of language use, and the range of knowledge, skills or abilities that underpin the test” (123-124). Teacher *arguments* (Tsagari & Cheng 2016) can help determine the role of LAL in test use and consequences.

The challenging nature of reading precisely justifies the investigation of LAL in the assessment of this skill. Despite the absence of a comprehensive theory of reading, recent L2 reading research widely agrees that the orchestration of complex reading processes makes the measurement of this skill particularly demanding while maintaining that the main goal of reading assessment is to collect evidence on “the skills, processes, and knowledge resources” (Grabe 2009, 353). Supposedly related to a curriculum in a specific educational context like the one in the current study, LAL would enable teachers to operationalize the precise reading construct underpinning the formal tool they developed for measuring L2 reading achievement and related subskills (Alderson 2000) or strategies (Phakiti 2003). Their arguments should support an accurate understanding of the nature of reading and reading test design (Fulcher & Davidson 2009).

5. Context of the study

5.1 Characteristics of the educational setting

Research on teaching and assessing English as a foreign language in Tunisian higher education has revealed a number of issues. Tunisian departments of English, in particular, have generally been faced with the challenge of offering a university degree in English within a country that is still influenced by the French educational system (Al-Khatib 2008; Battenburg 2006). Studies have shown that almost half of the teaching staff are “non-standard” (Abdesslam 2011, 140) and sometimes comprise “unqualified teachers” (Labassi 2009, 249). In terms of background, these teachers generally receive no pre-service academic training in language testing (Maaoui 2016) and many “item writers overlook test specs” (Hidri 2016, 34). Besides, while European educators and test designers have guidelines to help them identify the set of abilities to measure through tests (see the Common European Framework of Reference–CEFR, Council of Europe 2001), the Tunisian higher education system lacks a standardization of descriptors of English language ability. Despite the recent proliferation of language assessment research in the country (Hidri 2012, 2016; Maaoui 2016) limited research addressed LAL in classroom assessment in relation to writing assessment (Athimni 2018). Research on LAL can play a role in determining and enhancing language assessment quality.

5.2 Research questions

To address the issue of LAL versus validity, three main research questions were developed while accounting for personal, educational and sociocultural factors interacting with LAL. The questions that drove this study were as follows:

(1) What do Tunisian higher education teachers of English know about language assessment in general and reading comprehension measurement in particular?

(2) What do these teachers report about the reading test design process?

(3) What kind of validity evidence does the reported test design provide in this context?

This study would contribute to determining the quality of teacher-led assessment in Tunisian higher education, revealing the potential dangers of uninformed high-stakes test use in a testing-oriented educational context (Hidri 2012, 2015; Maaoui 2016).

6. Research Design

The study framework discussed earlier served to collect and analyze data from semi-structured interviews with three tertiary-level English as a foreign language teachers who were also the item writers and scorers of a reading test. This was a mid-term exam for all first-year students in this department in this Tunisian University but for reasons of anonymity, no further information can be disclosed.

The researchers were granted approval to collect data from these assessors who were submitted to a semi-structured or “standard” (Merriam 1998) 30-minute interview session based on a series of questions. As displayed in Table 1, the interview questions revolved around the following aspects: (1) the item writers’ profiles; (2) knowledge of language assessment principles; (3) knowledge of assessment design; and (4) skills in designing the reading test.

Table 1: ***The interview guiding questions***

Information elicited	Guiding questions
(1) The item writers' profiles	- Gender - Educational background - Teaching experience
(2) Knowledge of language assessment principles	- Background in language assessment - The way they learned to use assessment in their classrooms - Whether they ever worked with standardized tests and what they think of them - Knowledge of how to design formal tests
(3) Knowledge of assessment design	- Resources, guides, theoretical or assessment framework used to design this high-stakes test
(4) Skills in designing the reading test design	- Construct definition - The alignment between the construct and test method - The agreement on performance criteria by the three item writers

The researchers probed and prompted the test designers to report on their test design and intentions thus helping those with little assessment meta-language to clarify, rephrase or explain their ideas. The interviews were recorded with the informants' approval. The three informants were anonymized using pseudonyms for ethical purposes (see also Table 3). The immediate transcriptions allowed a preliminary analysis and comparison of the responses (Block 2000) after piloting with two university teachers in other departments of English that led to the simplification of the interview questions to maximize data validity. Given the small size of the participants, additional face-to-face meetings and phone conversations were required to clarify ambiguities for a more accurate portrayal of the participants' points of view (Marshall, Cardon, Poddar, & Fontenot 2013). The transcripts and summary notes were analyzed immediately after each session by summarizing, interpreting, comparing, and categorizing what the study participants verbalized about their *knowledge* and related *skills* in crafting the reading achievement test (Merriam, 1998). In line with the interview topics outlined above, coding was also used to label meaningful segments of the transcribed text into categories (Lodico, Spaulding, & Voegtle 2010) to identify emerging themes in relation to the item writers' profiles, assessment knowledge and skills.

7. The Paper-and-pencil Reading Comprehension Test Design

For accountability purposes, the three item writers who were equally teachers of reading and writing courses co-designed this high-stakes test (see the Appendix) to measure 571 students' reading achievement in first-year English at university. In the department of English, where the study was carried out, the students sat for a language exam based on the measurement of only two language skills: reading and grammar. They were given two hours to complete the two tests.

This reading course to which this achievement test is related has a number of characteristics of relevance to this study. The syllabus included objectives like the specific reading strategies of skimming and scanning, understanding reference words, using the context to understand new word meaning, summarizing the text ideas, inferencing and paraphrasing. Additionally, the reading passages slightly varied in length (between 350 and 600 words) but all relied on authentic texts of various genres (literary, press ..., etc.). The reading tasks were generally developed or adapted by the teachers of the course. It is important to note that the structure of the students' worksheets containing the text and questions resembled the reading test in terms of the question types and format. It follows that the reading achievement test should tap into the same strategies taught during the reading classes (Bachman & Palmer 1996).

The reading test was scored by the same item writers. The scores <10 were interpreted as "fail" while scores ≥10 were interpreted as "pass" on a scale from 0 to 20. A summary of the scores after the administration of this mid-term test shows a variance of 10.20, thus reflecting a problem possibly related to the test takers' performance if compared to the mean (7.69). Standard deviation (3.19) also indicates that the scores are skewed to the left, displaying many low scores, which might be the result of either a variance in the test takers' reading ability or other factors that could have affected test design and/or performance. In any case, these results could be indicative of either a low reading ability or issues of reading instruction for the different test stakeholders.

As displayed in Table 2, the test consisted of five parts comprising different item types. The first part included three short-answer questions. In the second section, the test takers responded to true/false statements with a justification (Alderson 2000). This section combined fixed and constructed-response items. The third section engaged the test takers in paraphrasing two statements and producing a short summary of specific text ideas in the fourth part. The last question required the test takers to find the equivalents of four words/expressions provided on their answer sheets.

Table 2: Description of question types and items in the reading comprehension test

	Question type	Item format	Number of items
Section 1	Constructed response/short answer	Open-ended format	3
Section 2	True/false with justification	Fixed/open-ended format	3
Section 3	Understanding and paraphrasing statements	Open-ended/constructed format	2
Section 4	Summary of specific text ideas	Open-ended/constructed format	1
Section 5	Vocabulary	Open-ended format	4

The test was based on a text of 532 words adapted from an American weekly news magazine (see Appendix A).

8. Results and Discussion

8.1 The teachers' profiles

The three teachers' profiles, academic backgrounds, and experience are summarized in Table 3. The interviewees were a female novice teacher with one year of teaching experience and two male participants with nine years of teaching experience each. These teachers were regular teachers of reading and writing courses to first-year students in the Department of English in the Tunisian university under study.

Table 3: The Teachers' Background

Name	M/ F	Level and field of study	Experience in teaching reading	Background in language testing
Sam	Male	MA in English linguistics	4 years	Little experience in test design from his experience as a former teacher in secondary education
Neil	Male	MA in English literature	2 years	Very little experience in test design
Haidy	Female	BA in English literature	1 year	Theoretical knowledge of basic principles of testing acquired through a language testing course

8.2 The teachers' assessment knowledge base

This specific aspect of assessment knowledge addressed the know-what and the know-how of language assessment. The results indicated that the test designers had diverging assessment backgrounds. For instance, Haidy exceptionally had some theoretical knowledge of language testing. She reported that she had restricted training on "basic testing principles" related mainly to "techniques and question types". Besides, only Sam was a linguistics major who informally communicated to the researchers that he had about five years of teaching experience in secondary education through which he had "learnt how to design a test from more experienced teachers" and "the teacher trainer". This implied that he had mainly some test design, administration and marking *hands-on experience*. This may confirm previous research findings about Tunisian teachers' development of limited assessment competencies "through regular practice of assessment tasks" (Athimni 2018, 184).

When prompted about their knowledge of the difference between reading comprehension in testing and non-testing situations (Bachman & Palmer 2010), the three teachers had more or less the same perceptions. Sam demonstrated an awareness that learners need to learn *how* to read. However, he considered reading test performance to be slightly *different* from reading in both learning and real-life contexts. He explained that, in a testing situation, the students should learn how to employ their reading strategies to demonstrate their understanding of a text. For Neil, the test taker is solely “obsessed with accurate answers”, thus, implying that for Tunisian students a test should mainly lead to a score. Haidy confirmed Sam’s *logic* by highlighting a difference between “the way we read in learning and during test-taking”. The teachers’ statements reflected a conception of the test as a problem-solving situation (Maaoui 2016) whereby test takers are mainly concerned about the final “score”. In a nutshell, the teachers were more preoccupied with performance *outcomes* and the product of formal testing. They failed to demonstrate accurate knowledge of what “test performance” entails in terms of predicting language performance in the target language use domain (Bachman & Palmer 2010).

With the limited opportunities for assessment knowledge building in the local context, teachers tended to rely on alternative methods to cope with the demands of *formal testing* as a priority. What generally informed the teachers’ work as assessors can be summarized in a complex pattern identified in the data. One informant indicated that his test development process “followed the pattern of standardized tests of English like the TOEFL exam”. Nevertheless, all of them agreed about the inevitability of “following the design of reading comprehension tests used by previous teachers of reading comprehension courses” in local higher education institutions. The teachers described this as common practice that developed into a real tradition within this teaching community. In the light of these findings, assessment knowledge can be described as collaborative, implicit and limited. With the absence of “assessment frameworks” (Fulcher & Davidson 2009) and clear assessment guidelines, teachers were found to compensate for the knowledge gap with tacit assessment norms. Most importantly, this generally reflected insufficient knowledge of the field of language assessment.

8.3 The teachers’ interpretation of the reading test design process

The study of LAL in this particular context had to consider key stages of test design including the definition of the reading construct. When asked about what they intended to measure, the interviewees broadly described the tool used as “a reading test”. It was no surprise that the test had the overt *purpose* (Fulcher & Davison 2009) of examining the test takers’ reading *achievement* as stated by these teachers and required by the educational system. However, it is imperative to clarify that the teachers demonstrated no awareness of the need for test specifications. Therefore, it was necessary to further prompt the interviewees to explain this test design process and the *knowledge* underlying their decision-making process.

Despite the absence of specs, the item writers believed that they "planned" (Sam) and "drafted" (Haidy) questions reflecting "adequate" and "varied" testing techniques. Their interpretation of this test planning phase showed a chief concern with what they referred to as "questioning techniques" or "questions" and "text type". What was remarkable here was the ambiguity divulged through the test designers' choice of the test item formats as they declared having relied on "different techniques" of open-ended, true/false, summarizing and vocabulary questions to allow their students to "show their comprehension" believing that closed items are limited in this respect. However, they justified their use of "open ended questions" mainly as a way to avoid cheating and to assess the learners' understanding of the text "in a good way" (Haidy). The interview revealed that the planning stage did not involve clear scoring procedures for each item. This might be related to the absence of clear performance criteria and understanding of the nature of reading as a skill (Alderson 2000; Grabe 2009). In the case of constructed-response questions, the three test designers merely provided a broad description of possibly "acceptable" answers with no real criteria for evaluating or rating a wide range of potential responses that may affect inter- and intra-rater reliability (Bachman & Palmer 2010).

The interviewees' repeated arguments supported many of the assessment research findings in this educational context. These teachers generally used "testing" and "assessment" interchangeably and tended to refer mainly to their responsibility in formal "testing" and the measurement of learning outcomes. Within the same vein, they reported an intentional *familiarization* with achievement test item formats during reading instruction. This was typical in an educational landscape predominantly marked by a concern with summative assessment. In this culture where teaching to the test prevails (Athimni 2018; Hidri 2016), these teachers tended to favour specific test item formats just to minimize cheating risks rather than to reach their *intended* testing purpose (Fulcher 2017). Doing so, they totally ignored any risk of low validity (Alderson 2000; Fulcher 2010, 2017). Accordingly, they based their arguments on the *practice* of relying on "techniques" just because of their normalized use. When further prompted about their knowledge of these "commonly used techniques", Sam declared having seen them "in tests given by other teachers" further supporting earlier interpretations of the role the educational community plays in shaping LAL and assessment practices (Xu & Brown 2016).

The test designers' explanation of the reading test design process *per se* revealed rather diverging claims. Haidy noted that two factors were significant for developing good and "clear" tests. First, she raised the issue of time. She reported that she felt that her teaching workload and commitments prevented her from developing a "convincing test". She added that the absence of any framework guiding language teachers in test design was a second obstacle. She admitted that within her limited training in language testing, she became "aware" of the degree to which language test design is "a complicated process".

The interviewees explained that questions had to be in "parallel" (Neil and Sam) with the techniques with which the test takers were familiar in class. This visibly reflected the teachers' ignorance of objective or scientific criteria for the selection of reading texts, which might have led to *uninformed decisions* during test design and scoring (Athimni 2018; Maaoui 2016). Even though Haidy's interpretation of the test design and the logic underlying it reflected doubts about the quality of this test, the teachers' overall discourse was marked by a total absence of clear plans, actions or evaluation of such actions prior to test item development or administration. This revealed an uncertainty and a lack of agreement on *what* was measured and *how* due to their ignorance of testing principles. The teachers' responses also reflected these stakeholders' inability to interpret their logic underlying the test design because of their "assessment illiteracy" (Popham 2004). This somewhat varying deficiency was found at the level of both *knowledge* and *skills.* The absence of test specs might have affected validity in this high-stakes testing context.

The question of text selection was equally addressed in the interview. Both Neil and Sam declared that the most demanding task was "careful" text selection. Sam particularly emphasized the difficulty of selecting a text of a particularly appropriate length and difficulty level or "up to their level" as Sam put it. The main criterion in text selection was the "familiarity" with the topic. To be able to align the text difficulty to the learners' level, these item writers reported that they simply knew that the first year of higher education corresponded to "above an intermediate level" (Sam). They unanimously revealed that it was through teaching and their weekly interaction with the students that they decided about the students' level of average reading ability. When the researchers prompted the informants to describe the test takers' level in terms of any "ability descriptors", the three participants failed to provide a thorough argument. It was evident that test use was not based on clear *criteria* relating to aspects of good performance as discussed in the literature (Alderson 2000; Grabe 2009).

The teachers' discourse reflected a broad concern about how they sought to lead their students to provide "correct responses". The informants also emphasized the pertinence of "clear" test questions so that "they can use appropriate strategies they learnt in class" (Haidy). According to the three teachers, students can prove that they have understood the text by answering the questions in the way required. Surprisingly, the three informants' description of their logic in predicting the test takers' processing of the test items in terms of strategic application was not in line with validity demands (Moss 2015). The three test designers considered the responses to the test questions as "obvious". This divulged the logic behind the absence of anticipated patterns of performance that would guide the subsequent rating of the students' answers. They explained that good students usually deal with the test questions strategically and avoid wasting time. Haidy added that good achievers often select the part of the text that is relevant to the answer. As to the purpose of these speculative predictions, the three teachers were confident about the usefulness of this stage before administering the test. However, they stated that they wrote an "approximate correction" of the test item to "predict any possible ambiguities or difficulties" during test taking (Sam). They mainly produced sample answers for the open-ended questions without really accounting for possible variations in the test takers' written responses.

To gain a better understanding of the teachers' levels of LAL, we provide a summary of the test designers' views of what each test item was intended to measure (Table 4). Regarding the first section of the test, Haidy explained that in items 1, 2 and 3 of the test, the students were meant to "use strategies that divulge their mastery of the course" while insisting on skimming and scanning as main strategies. For the same items, the second teacher (Neil) stated that it aimed at "understanding discourse, context, surrounding ideas and structures, extension of meaning using background knowledge" while being able to "paraphrase using underlining and note-taking" through "non-linear-reading". The third teacher (Sam) briefly stated that the test takers were led to paraphrase the text ideas. The three teachers' opinions clearly diverged as to what the item measured exactly. This generally reflected a lack of consensus regarding the construct that the item aimed to assess.

Table 4: The test takers' intended strategy use as perceived by the test designers

Test tasks/skills	Haidy	Neil	Sam
Open-ended/short answer	Skimming and scanning	Understanding discourse/non-linear reading	Paraphrasing the text ideas
True/false with justification	Being able to read and understand detailed ideas	Being able to justify what is correct and understand that an idea is false	Making the difference between correct and incorrect information
Paraphrasing statements	Getting the text meaning from the context Understanding keywords and finding equivalents to them	Expressing the original text ideas in a different way	Paraphrasing by using their words and by changing the sentence structure
Summary of specific text ideas	Summarizing the text ideas	Providing a coherent and well written summary	Summarizing the text using their own words
Vocabulary	Understanding the difficult vocabulary items of a text using the context	Guessing word meaning using the context	Understanding new words using the context

The second set of items (true/false statements) was intended by the test designers to measure the test takers' ability to differentiate between correct and incorrect statements and find the appropriate justifications in the text. The three participants' views about what these particular items aimed to assess were quite clear and uncontroversial. The third and fifth sections of the test were particularly problematic. The designers' aim behind the items in section three was broadly expressed as "getting the text meaning from the context [and] understanding keywords and finding equivalents to them" (Haidy). For Neil, the students had to express the original text ideas in a different way relying on "paraphrasing" as a key strategy. The item planning process reveals that the construct measured did not exclusively include reading *per se*.

Contrary to validity requirements, collecting evidence about the test takers' ability to comprehend a text in English was particularly problematic in this context. Several LAL and contextual factors might have led to an absence of a clear construct definition. The teachers' views exhibited limited knowledge of *test design principles*, most likely because of this lack of *professional development* opportunities or willingness on the part of teachers as admitted by the informants. Indeed, two of the teachers were literature majors (Haidy and Neil) and all the teachers had rather varying levels of "assessment illiteracy" (Popham 2004). At a macro level, this educational context might have been operating with normalized assessment practices that diverged from professional, ethical and expert demands. The teachers' discourse, in particular, echoed inconsistent views and beliefs, and most importantly, an absence of adequate assessment *knowledge*.

Despite the compromise among the three informants about their intention to measure the students' ability "to use appropriate reading strategies" (Sam and Nigel) in terms of what they are taught in class, there was a mismatch between the different intentions in using the different test items and performance criteria. These reported views were not necessarily only influenced by the teachers' academic training but also the dominant assessment culture. The teachers' *uninformed* test use was characterized by (1) a lack of theoretical knowledge about language testing principles and design, and (2) an *insufficient* awareness of the theoretical underpinnings of reading comprehension assessment. For instance, only one of the test designers acknowledged the test taker background and topical knowledge as important factors to account for in reading assessment. These findings are consistent with previous research on teacher-based assessment (Brindley 2001; Davison 2004).

9. Summary of the Results

The present study set up to investigate the LAL of three Tunisian tertiary education teachers who were involved in the design and scoring of an institutional mid-term exam supposedly related to a national reading curriculum. Interviewing teachers as key agents in classroom assessment in general and high-stakes testing in particular has proven to be useful in disclosing different LAL levels. Thus, LAL, which was found to be inadequate for these teachers, can be described in terms of degree, awareness, development and collaboration. Given this nature, it is possible to foster LAL through a self- or other-initiated planned teacher professionalization process.

The teachers’ assessment knowledge and beliefs might have had *unintended* consequences on reading ability measurement (Kane 2013). With LAL training and development as a relatively neglected professional area (Popham 2011), language teachers might simply rely on available sociocultural resources like habits or common practice to use assessment including the design and use of their tests. They can even be misled and draw wrong inferences about ability from the scores obtained without being aware of the impact this might have on the quality of measurement and on test takers’ lives. These issues further underscore the overlap between the *evidential* and the *consequential* basis of validity as a professional and ethical paradigm.

Putting the present findings in their educational, cultural and political contexts has required the elucidation of the possible factors that might have come into play in this specific assessment situation. Language test development in this tertiary institution does not differ much from reading instruction. The piloting of the interviews had revealed that high-stakes reading tests used in English departments for first-year English as a foreign language students were very similar to those used for instructional purposes. These tests never relied on multiple-choice questions. The empirical evidence discussed above supports previous research about LAL multidimensionality. LAL issues can also relate to gaps in theoretical knowledge, to misconceptions, a low quality of classrooms in terms of valid assessment. Hence, teacher-led assessment practices can be characterized by sketchy, intuitive and unprofessional practices.

Despite the importance of an overall *test purpose* rather dictated by the educational system, predictions about what would constitute reading achievement or lack of achievement were either too general or inexistent in this context. Teacher-based assessment deviated from expert principles and was implemented with almost a total absence of criteria, standards or guidelines in a global educational assessment scene that has sought to establish clear statements about students’ *intended achievements* in languages in terms of benchmarks and competencies (ILTA 2007).

The use of interview data to explore teachers' LAL vs. the *a priori* arguments about the nature of the reading constructs revealed a number of issues. The teachers' discourse clearly reflected an absence of consensus on the traits measured. This was due to poor test planning because of an ignorance of the demands of construct definition. Problems in LAL somehow resulted in an absence of test specifications or “blueprints” (Davidson & Lynch 2002; Fulcher 2017; Fulcher & Davidson 2009) in a system where teachers know little to nothing about the demands of valid measurement of language ability. The process of test design described by these teachers confirms that when LAL is deficient the remaining measurement process can be affected (Fulcher 2010). The study revealed the teachers’ direct writing of the test items without any written draft, consensus, discussion, testing and re-writing of the items as required in the language test-design literature (viz. Fulcher 2010, 2017). Therefore, issues of diverged views on *what* was assessed and *how* emerged in the data. This certainly pinpoints the value of investigating *a priori* arguments in test design.

With their lack of assessment expertise (Athimni 2018), the teachers might have ended up measuring a construct that was ill-defined. The deficiency in LAL might also have harmful effects on individual test takers, an erroneous interpretation of their language proficiency and wrong pedagogical decisions taken by the teachers and their institutions. Therefore, an intervention to develop teachers’ LAL could address theoretical models of languages (Fulcher & Davidson 2009) as well as the nature of reading as a skill and its sub-skills in addition to how they can be measured through test tasks. Such testing practices can be simply described as lacking a firm foundation that assured a purposeful and objective measurement of language ability. The vagueness in task specification might have led to vague and inaccurate measurement (Bachman 2002).

As revealed by the three teachers, the rationale for this achievement test was to measure the degree to which the students developed the reading strategies they had learnt in class (Sam). They showed a common concern about the lack of guidelines for how to assess learners’ reading ability and language test design in general. The present findings should be related to the subsequent discussion of the remaining study data to examine whether the way the test was designed could affect test quality in general and test validity in particular. Having accounted for the test designer helped to unveil fundamental aspects of direct relevance to validity. Because of the complexity of aligning assessment to the curriculum principles and LAL issues, high-stakes, classroom-based testing might affect validity in multiple ways. The absence of assessment standards, frameworks and professional development may hinder accurate test use (Bachman & Palmer 2010), muddy test scores, and lead to validity chaos (Fulcher 2010).

10. Limitations, Implications and Recommendations

The study has provided some significant context-specific findings comparable to previous assessment research in other educational settings. This piece of research sought to elucidate the relationship between test-related and non-test-related variables like test design in relation to validity. These findings raise issues of the validity of teacher-led assessment in the local context. These teachers were shown to be sometimes unprepared for the design of useful ability-measurement tools. The lack of agreement among the test designers can particularly be taken as an indication of a LAL deficit leading to their dependence on common practice rather than language assessment problems. This issue might have been intensified by contextual factors like the dominant assessment culture and the absence of assessment standards and guidelines.

It can be observed that language assessment implementation varies only slightly between Tunisian teachers and institutions in the tertiary level. Therefore, the study results may be said to have external confidence and can be compared across institutions. These findings indicated that interpreting LAL data while accounting for contextual factors could be more insightful than merely relying on a context-free, pre-established set of factors that guides LAL data elicitation. This study further confirmed the value of qualitative, in-depth research into the LAL nature and assessment consequences. Future research should address the implementation of assessment and the *logic* behind it as an integral LAL component. Interpretations of LAL should cater for interpersonal differences and trends within a community of practice. They should also triangulate data sources to check the teachers' claims and statements.

LAL research might also need to consider other stakeholders as an invaluable source of information. Such data can cover LAL vs. attitudes, conceptions of theoretical models of language, and definition of language ability. This new line of research should call for more contextualized and continued professional teacher development programs and courses combined with a clear assessment policy that recognizes the key role of the teacher. Accordingly, bringing classroom assessment to the fore has become an individual, institutional and national responsibility. Teachers themselves should give importance to their professional development to have more assessment expertise. Finally, teachers should also develop more awareness of their professional responsibility in assessing their learners in a valid way.

In summary, these findings have served to shed light on some issues in teacher-led assessment in Tunisian higher education. Assessment illiteracy and the unintentional *misuse* of high-stakes ability measurement may be attributed to the marginal value of assessment in teacher professionalization in Tunisia. This gap in LAL and inadequate test design might have led to wrong inferences about the ability measured and a compromising not only of the test in question but also of the test takers' academic future in this specific context.

References

Abdessalem, Tahar. 2011. "Scope, Relevance and Challenges of Financing Higher Education: The Case of Tunisia." *Prospects* 41:135-155.

Alderson, J. Charles. 2000. *Assessing Reading*. Cambridge: Cambridge University Press.

Allal, Linda. 2013. "Teachers' Professional Judgement in Assessment: A Cognitive Act and a Socially Situated Practice". In *Assessment in Education: Principles, Policy & Practice*, edited by Claire Wyatt-Smith, and Valentina Klenowski, 20(1):20-34.

Allal, Linda, and Lucie Mottier Lopez. 2014. "Teachers' Professional Judgment in the Context of Collaborative Assessment Practice". In *Designing Assessment for Quality Learning*, edited by Claire Wyatt-Smith, Valentina Klenowski, and Peta Colbert, 151-165. Netherlands: Springer.

Athimni, Moez. 2018. "Investigating Assessment Literacy in Tunisia: The Case of EFL University Writing Teachers". In *Assessing EFL Writing in the 21st Century Arab World*, 163-189. Cham: Palgrave Macmillan.

Bachman, Lyle F. 2002. "Some Reflections on Task-based Language Performance Assessment". *Language Testing,* 19:453-476.

Bachman, Lyle F., and Adrian S. Palmer. 1996. *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford: Oxford University Press.

Battenburg, John D. 2006. "English Language Teaching in Tunisia", 5-12. Retrieved from http://cla.calpoly.edu/_ jbattenb/world/tunisia.html.

Black, Paul, and Dylan William. 2010. "Inside the Black Box: Raising Standards through Classroom Assessment." *Phi Delta Kappan* 92(1):81-90.

Block, David. 2000. "Problematizing Interview Data: Voices in the Mind's Machine?" *TESOL Quarterly* 34(4):757-763.

Brindley, Geoff. 2001. "Outcomes-based Assessment in Practice: Some Examples and Emerging Insights." *Language Testing* 18(4):393-407.

Brown, Galvin T. L. 2006. “Teachers' Conceptions of Assessment: Validation of an Abridged Version.” *Psychological Reports* 99(1):166-170.

—. 2008. “Assessment Literacy Training and Teachers’ Conceptions of Assessment.” In *Thinking about Teaching and Learning*, edited by Christine M. Rubie-Davies, and Catherine Rawlinson, 285-302. New York, NY: Nova Science.

—. Stephen Irving, and Peter Keegan. 2007. *An Introduction to Educational Assessment, Measurement and Evaluation: Improving the Quality of Teacher-based Assessment*. New Zealand: Pearson Education.

—. Kerry J. Kennedy, Ping Kwan Fok, Jacqueline Kin Sang Chan and Wai Ming Yu. 2009. “Assessment for Student Improvement: Understanding Hong Kong Teachers’ Conceptions and Practices of Assessment.” *Assessment in Education: Principles, Policy & Practice* 16(3):347-363.

Constantinou, Filio, Victoria Crisp, and Martin Johnson. 2017. “Multiple Voices in Tests: towards a Macro Theory of Test Writing.” *Cambridge Journal of Education*: 1-16.

Davidson, Fred, and Glenn Fulcher. 2007. “The Common European Framework of Reference (CEFR) and the Design of Language Tests: A Matter of Effect.” *Language Teaching* 40:231-241.

Davidson, Fred, and Brian K. Lynch. 2002. *Testcraft: A Teacher’s Guide to Writing and Using Language Test Specifications.* New Haven, CT: Yale.

Davison, Chris. 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.

—. and Constant Leung. 2009. “Current Issues in English Language Teacher-based Assessment.” *TESOL Quarterly* 43(3):393-415.

DeLuca, C., and A. Bellara. 2013. “The Current State of Assessment Education: Aligning Policy, Standards, and Teacher Education Curriculum.” *Journal of Teacher Education* 64(4):356-372.

DeLuca, Christopher, and Aarti Bellara. 2017. “Developing Assessment Capable Teachers in this Age of Accountability.” *Assessment in Education: Principles, Policy & Practice* 24(2):121-126.

Deneen, Christopher Charles, and Gavin T. L. Brown. 2016. “The Impact of Conceptions of Assessment on Assessment Literacy in a Teacher Education Program.” *Cogent Education* 3(1): 1225380: 1-14.

Falsgraf, Carl. 2005. “Why a National Assessment Summit? New Visions in Action.” *National Assessment Summit*. Meeting conducted in Alexandria, Va.

Fulcher, Glenn. 2017. “The Practice of Language Assessment”. In *Handbook of Research in Second Language Teaching and Learning*, edited by Eli Hinkel, 463 - 475. London and New York: Routledge.

—. 2012. “Assessment Literacy for the Language Classroom.” *Language Assessment Quarterly* 9(2):113-132.

—. 2010. *Practical Language Testing*. Routledge.

—. and Fred Davidson. 2009. “Test Architecture, Test Retrofit.” *Language Testing* 26(1):123-144.

Fulmer, Gavin W., Iris C. H. Lee, and Kelvin H. K. Tan. 2015. “Multi-level Model of Contextual Factors and Teachers’ Assessment Practices: An Integrative Review of Research.” *Assessment in Education: Principles, Policy & Practice* 22(4):475-494.

Grabe, William. 2009. *Reading in a Second Language: Moving from Theory to Practice*. Cambridge University Press.

Harding, Luke, and Benjamin Kremmel. 2016. “Teacher Assessment Literacy and Professional Development.” In *Handbook of Second Language Assessment*, edited by Dina Tsagari, and Jayanti Banerjee, 413-427. Berlin/New York: Mouton De Gruyter.

Hasselgreen, Angela. 2008. “Literacy in Classroom Assessment (CA): What Does This Involve?” In the *5th Annual Conference of EALTA*. Athens, Greece. Retrieved from
http://www.ealta.eu.
org/conference/2008/docs/sunday/panel/Literacy%20in%
20classroom%20assessment. pdf.

—. Cecilie Carlsen, and Hildegunn Helness. 2004. *European Survey of Language Testing and Assessment Needs. Part One: General Findings*. Retrieved from http://www.ealta.eu.org/resources.htm.

Hidri, Sahbi. 2012. “Assessing Static vs. Dynamic Listening: Validation of the Test Specifications.” Unpublished PhD thesis. ISLT: University of Carthage.

—. 2016. “Conceptions of Assessment: Investigating what Assessment Means to Secondary and University Teachers.” *Arab Journal of Applied Linguistics* 1(1): 19-43. Retrieved from
http://www.arjals.com/ojs/index.php/Arjals2016/article/view/43.

Hill, Kathryn. 2017. “Understanding Classroom-based Assessment Practices: a Precondition for Teacher Assessment Literacy.” *Papers in Language Testing and Assessment* 6(1): 1-17. Retrieved from http://arts.unimelb.edu.au/__data/assets/pdf_file/0007/2349862/6_1_SI 1Hill.pdf.

Hill, Mary, Bronwen Cowie, Alison Gilmore, and Lisa F. Smith. 2010. "Preparing Assessment-capable Teachers: What Should Pre-service Teachers Know and be Able to Do?" *Assessment Matters* 2: 44e64. Retrieved from http://search.informit.com.au/documentSummary;dn=330451040688063;res=IELHSS.

Huhta, Ari, Hirvelä, Tuija, and Banerjee, Jayanti. 2005. "European Survey of Language Testing and Assessment Needs: Report: Part Two– Regional Findings". Retrieved from http://www.ealta.eu.org/documents/resources/survey-report-pt1.pdf.

International Language Testing Association. 2007. "ILTA Guidelines for Practice."

Inbar-Lourie, Ofra. 2008. "Constructing a Language Assessment Knowledge Base: A Focus on Language Assessment Courses." *Language Testing* 25(3):385-402.

—. 2016. "Language Assessment Literacy." In *Language Testing and Assessment*, edited by Elana Shohamy, Lair Or, and Stephen May, 1-14. Springer International Publishing.

—. 2013. "Guest Editorial to the Special Issue on Language Assessment Literacy." *Language Testing* 30(3): 301-307.

Kane, Michael T. 2006. "Validation." In *Educational Measurement* (4th ed.), edited by Robert L. Brennan, 17-64. Washington, DC: The National Council on Measurement in Education & the American Council on Education.

—. 2013. "Validating the Interpretations and Uses of Test Scores." *Journal of Educational Measurement* 50(1): 1-73.

Labassi, Tahar. 2009. "Periphery Non-Anglophone Scholarship in English-only Journals: Conditions of a Better Visibility." *Changing English* 16(2):247-254.

Lodico, Marguerite G., Dean T. Spaulding, and Katherine H. Voegtle. 2010. *Methods in Educational Research: From Theory to Practice* (Vol. 28). John Wiley & Sons.

Looney, Anne, Cumming, Joy, Fabienne van Der Kleij, and Karen Harris. 2017. "Reconceptualising the Role of Teachers as Assessors: Teacher Assessment Identity." *Assessment in Education: Principles, Policy & Practice*, *25*(5):442-467.

López Mendoza, Alexis A., and Ricardo Bernal Arandia. 2009. "Language Testing in Colombia: A Call for more Teacher Education and Teacher

Training in Language Assessment." *Profile Issues in Teachers' Professional Development* 11(2):55-70. Retrieved from http://revistas.unal.edu.co/index.php/profile/article/view/11442.

Maaoui, Asma. 2016. "Test-taker Beliefs, Strategies and Achievement in EFL Reading". Unpublished PhD thesis. University of Carthage, Tunis, Tunisia.

Marshall, Bryan, Peter Cardon, Amit Poddar, and Renee Fontenot. 2013. "Does Sample Size Matter in Qualitative Research?: A Review of Qualitative Interviews in IS Research." *Journal of Computer Information Systems* 54(1):11-22.

Merriam, Sharan B. 1998. *Qualitative Research and Case Study Applications in Education*. San Francisco, CA: Jossey-Bass.

Miller, M. David, Robert L. Linn, and Norman E. Gronlund. 2009. *Measurement and Assessment in Teaching* (10th ed.). Upper Saddle River, NJ: Pearson Education.

Moss, Pamela A. 2015. "Shifting the Focus of Validity for Test Use." *Assessment in Education: Principles, Policy & Practice* 23(2):236-251.

Newton, Paul E. 2016. "Macro- and Micro-validation: Beyond the 'Five Sources' Framework for Classifying Validation Evidence and Analysis." *Practical Assessment, Research & Evaluation*: 21(12): 1-13. Retrieved from http://pareonline.net/getvn.asp?v=21&n=12.

Phakiti, Aek. 2003. "A Closer Look at the Relationship of Cognitive and Metacognitive Strategy Use to EFL Reading Achievement Test Performance." *Language Testing* 20(1): 26-56.

Pill, John, and Luke Harding. 2013. "Defining the Language Assessment Literacy Gap: Evidence from a Parliamentary Inquiry." *Language Testing* 30(3):381-402.

Popham, W. James. 2009. "Assessment Literacy for Teachers: Faddish or Fundamental?" *Theory into Practice* 48(1):4-11.

—. 2011. "Assessment Literacy Overlooked: A Teacher Educator's Confession." *The Teacher Educator* 46(4):265-273.

—. 2014. "Looking at Assessment through Learning-colored Lenses." In *Designing Assessment for Quality Learning*, edited by Claire Wyatt-Smith, Valentina Klenowski, and Peta Colbert, 183-194. Netherlands: Springer.

—. 2004. "Why Assessment Illiteracy is Professional Suicide." *Educational Leadership* 62(1):82-83.

Rea-Dickins, Pauline. 2004. "Understanding Teachers as Agents of Assessment." *Language Testing* 21(3):249-258.

—. 2007. “Classroom-based Assessment: Possibilities and Pitfalls.” In *International Handbook of English Language Teaching*, edited by Jim Cummins, and Chris Davison, 505-520. US: Springer.

Scarino, Angela. 2017. “Developing Assessment Literacy of Teachers of Languages: A Conceptual and Interpretive Challenge.” *Papers in Language Testing and Assessment* 6(1): 18-40. Retrieved from http://arts.unimelb.edu.au/__data/assets/pdf_file/0008/2349863/6_1_SI2Scarino.pdf.

—. 2013. “Language Assessment Literacy as Self-awareness: Understanding the Role of Interpretation in Assessment and in Teacher Learning.” *Language Testing* 30(3):309-327.

Sellan, Rajenthiran, and Assessment Board. 2017. “Developing Assessment Literacy in Singapore: How Teachers Broaden English Language Learning by Expanding Assessment Constructs.” *Papers in Language Testing and Assessment* 6(1):64-87. Retrieved from http://arts.unimelb.edu.au/__data/assets/pdf_file/0006/2349933/6_1_SI4Sellan.pdf.

Shohamy, Elana. 1998. “Inside the ‘Black Box’ of Classroom Language Tests.” *Studia Anglica Posnaniensia* 33:343-352.

Smith, Lisa. F., Mary F. Hill, Bronwen Cowie, and Alison Gilmore. 2014. “Preparing Teachers to Use the Enabling Power of Assessment.” In *Designing Assessment for Quality Learning*, edited by Claire Wyatt-Smith, Valentina Klenowski, and Peta Colbert, 303-323. Netherlands: Springer.

Stiggins, Rick. 2006. “Assessment for Learning: A Key to Motivation and Achievement.” *Edge* 2(2):3-19.

—. 1991. “Assessment Literacy.” *Phi Delta Kappan* 72(7):534-539.

—. 2004. “New Assessment Beliefs for a New School Mission.” *Phi Delta Kappan* 86(1):22-7.

—. 2001. “The Unfulfilled Promise of Classroom Assessment.” *Educational Measurement: Issues and Practice* 20(3):5-15.

Sultana, Nasreen. 2019. “Language Assessment Literacy: An Uncharted Area for the English Language Teachers in Bangladesh.” *Language Testing in Asia* 9(1):1-14.

Taylor, Linda. 2013. “Communicating the Theory, Practice and Principles of Language Testing to Test Stakeholders: Some Reflections.” *Language Testing* 30(3):403-412.

—. 2009. “Developing Assessment Literacy.” *Annual Review of Applied Linguistics* 29:21-36.

Tsagari, Dina. 2016. “Assessment Orientations of Primary State School EFL Teachers in Two Mediterranean Countries.” *CEPS Journal:*

Center for Educational Policy Studies Journal. Special thematic issue entitled *'Assessment of Young Foreign Language Learners'* 6(1):9-30.

—. and Jayanti Banerjee. 2015. "Language Assessment in the Educational Context." In *Handbook of Educational Linguistics*, edited by Martha Bigelow, and Johanna Ennser-Kananen, 339-352. New York: Routledge/Taylor & Francis Group.

—. and Liying Cheng. 2016. "Washback, Impact and Consequences." In *Encyclopedia of Language and Education* (3rd ed.), Volume 7: Language Testing and Assessment, edited by Elana Shohamy, and Nancy H. Hornberger, 1-13. Springer Educational Linguistics.

—. Tony Green, Karin Vogt, Ildiko Csepes, and Nicos Sifakis. 2017. "Our TALE: Developing Online Teacher Assessment Literacy Training Materials." Paper presented at the 39th Language Testing Research Colloquium, Universidad de Los Andes, Bogotá, Colombia. July 17-21, 2017 http://www.iltaonline.com/LTRC/.

—. and Karin Vogt. 2017. "Assessment Literacy of Foreign Language Teachers Around Europe: Research, Challenges and Future Prospects." *Papers in Language Testing and Assessment* 6(1):41-63. Special Issue edited by Kathryn Hill on '*Teacher Assessment Literacy in Second and Foreign Language Education.'*
http://arts.unimelb.edu.au/ltrc/research/publications/papers/current-issue.

Vogt, Karin, and Dina Tsagari. 2014. "Assessment Literacy of Foreign Language Teachers: Findings of a European Study." *Language Assessment Quarterly* 11(4):374-402.

Weir, Cyril. J. 2005. *Language Testing and Validation: An Evidence-based Approach.* Houndmills: Macmillan.

Xu, Yueting, and Gavin G. T. Brown. 2016. "Teacher Assessment Literacy in Practice: a Reconceptualization." *Teaching and Teacher Education* 58:149-162.

Yan, Jin. 2010. "The Place of Language Testing and Assessment in the Professional Preparation of Foreign Language Testers in China." *Language Testing* 27(4):555-584.

Appendix: Reading Comprehension Test

Preying on Children

Over the last decade, thousands of kids from all over the world have been smuggled into Europe to do dirty work. In September police arrested a couple in the southern Italian region of Calabria for allegedly buying an Albanian boy from a Dunes-run trafficking gang, one of a clutch of gangs involved in supplying Italians with Albanian kids for illegal adoptions. Eastern Europe's boys work in petty crime. Chinese children are trafficked to work in European sweatshops or on the streets. West African kids are sneaked into Europe to work as domestic servants. The 2001 discovery of a Nigerian boy's torso in the Thames led British police to arrest 21 suspected traffickers this summer.

Globalization, post-Soviet poverty and the European Union's newly porous borders have made child trafficking the world's fastest-growing branch of organized crime. Reliable statistics are scant, but the United Nations puts the worldwide number of children trafficked at 1.2 million a year. A 2001 EU study estimated that 120,000 women and children are clandestinely brought into Western Europe annually.

Police organizations take the problem seriously, but investigations are often handicapped by Europe's fragmented legal system. Laws on trafficking and migration vary enormously among the EU's 15 member countries. And with 10 new members set to join the Union–many of them, like Hungary, source countries for trafficked women and children there are sure to be new complications. Next summer could see a huge rise in Eastern European young people heading to Western Europe, where they'll be vulnerable to exploitation, fears Lars Loot of the Children's Unit at the council for Baltic States. "In a large Europe. it's naturally even easier for under-18s to go across borders without authorities knowing," he says.

Legal loopholes can leave victims unprotected and traffickers unpunished. Until recently the Greeks arrested child-trafficking victims. This year the US State Department threatened Greece with sanctions if it didn't improve its counter trafficking policies. The country has since promised a program of victim assistance. Italy has led Europe with progressive legislation on humanitarian visas to ensure that victims aren't repatriated and re-trafficked, and in August passed a tough law equating trafficking with slavery.

Worried that trafficking will only grow in the years ahead, Europe knows a more coordinated approach is needed. Last year's Brussels Declaration was the EU's first attempt to develop a comprehensive policy on trafficking that includes prevention, punishment and rehabilitation of the victims. Last week EU Justice and Home Affairs ministers agreed to grant limited-term visas to trafficked people who cooperate with police. Source countries, including Croatia and Hungary, have started counter trafficking campaigns. But success is a relative term.

Whether Albanian or Nigerian or Chinese, the traffickers prey on poverty and ignorance. "The trade has really grown recently, mainly because of the downward trend of economic development [in West Africa]," says UNICEF's Dakar-based spokesperson Margherita Amodeo. "Families that don't have anything are the most vulnerable." But other regional issues–a lack of education, the low status of girls, weak legal systems–also contribute to the problem. Vidomegon, a common West African practice in which families send children to live with wealthier relatives or friends to get a better education, is one way traffickers find to acquire kids.

By Carla Power
With Fiat Radcliffe in London & Karen Macgregor in Durban, South Africa
Newsweek November 17, 2003
(abridged)

Reading (text: Preying on Children)

Write all your answers on this sheet

I) Open-ended questions: Answer the following questions in your own words. (3 marks)

1.What are some of the reasons behind the illegal trafficking of children?
....

2. What were the major difficulties that faced police organizations?
....

3. How did the US State Department bring Greece under its control?
....

II) Write "True" or "False" in front of the following statements and justify your answers (6 marks)

1. The European Union is in control of its borders (…..)
....
....

2. Italy regards trafficking as bondage (…….)
....
....

3. Traffickers have a powerful grip on poor people (…….)
....
....

III) Explain the following statements (3 marks)

1. "Legal loopholes can leave victims unprotected and traffickers unpunished" (L. 22)
....
....
....

2. "**Whether** Albanian or Nigerian or Chinese, the traffickers prey on poverty and ignorance" (L. 35)
....
....
....

IV) Summary

Sum up briefly the effects of child trafficking on the European Union's stability (6 marks)

....
....
....
....

V) Find the equivalents to each of the following words: (2 marks)

- Trafficked = ...
- Trivial = ...
- Hardly enough = ...
- Secretly = ...

CHAPTER NINE

STUDENTS' VOICES ON THE UNSPOKEN ORACY CONSTRUCT: "FIND OUT HOW TO DO IT ON YOUR OWN!"

ANNE-GRETE KALDAHL

The aim of this chapter[1][2] is to gain insight into lower secondary students' perceptions of the assessment of oracy across subjects, and thereby capture the underlying oracy construct. This chapter draws on a rhetorical topos analysis of qualitative data generated through focus group interviews with tenth-grade students representing two lower secondary schools in Norway. The results indicate that the content of the students' utterances is the most valued part of the oracy construct. Their ability to display character and appeal to the teachers' sense of emotions are two other important dimensions. Students rely on their common and experience-based knowledge as a consequence of being exposed to teaching and feedback without a meta-language on oracy, hence being subjected to the reproduction of social inequalities.

1. Introduction

Politically the quality improvement of assessment practices in education is in focus (Mercer, Warwick and Ahmed 2017). The background for the research presented in this chapter is the comprehensive reforms in Norway (LK06) where basic competencies (oral competence (oracy), writing, reading, numeracy, and digital competence) are supposed to be taught and assessed across subjects (Knowledge Promotion 2007). Norwegian (L1) has the main responsibility for teaching oracy;

[1] In this study, and in alignment with the intentions of the reform and DeSeCo, the term *competence* substitutes the Norwegian term *ferdighet,* which is usually translated as "skills" as was done in Hertzberg & Roe (2016).

[2]This chapter is part of a mixed methods study with a parallel convergent design (Creswell 2014) based on three separate articles.

consequently, oracy competencies in L1 are expected to be adapted and practised in other subjects. The inclusion of oracy as a key competency in the Norwegian curriculum, is making all teachers language teachers, who are supposed to teach, assess oracy and arrange oral exams, within their subject disciplines. However, the assessment plan was not revised accordingly (Berge, Evensen and Thygesen 2016); consequently, leaving each individual teacher with the challenges of an assessment system that might not support the visions of the curriculum. This the autumn of 2020, a revised version of the LK06 curriculum based on these cross-disciplinary competencies, now called critical competencies, will be introduced on the grounds of deep-level processing as opposed to surface-level processing (Marton and Säljö 1976). The curriculum is developed through a cognitive and social cultural view of learning, where it should be acknowledged that language and oracy play essential roles in meaning making and knowledge processes, and that communities influence humans, who they belong to, and the cultural discourses they participate in (Vygotsky 1978, Wenger, McDermott, and Snyder 2002). The ability to create and express one's opinion word for word instead of repeating requires attention and represents deep learning, and the words are necessary in order to think and reason and develop intellectually (Wilkinson 1965, 13); thus, the ability to think in words and be able to express them should be given major attention in education. Thinking is therefore a competency intertwined in all these so-called critical competencies, and strongly linked to oracy and "oral composition as preliminary to written composition" (Wilkinson 1965, 11). For these reasons, the urgency arises to research the status quo for oracy in Norwegian schools. The ones who experience the teaching and assessment practices over time are the students. Therefore, the students' voices are important to listen to and include in research. The purpose of this chapter is to present new insights into how tenth-grade lower secondary students (aged 15-16) give meaning to oracy (the students' doxa) (Berge 1990), as well as the teaching and assessment (the qualifying norms) (Berge 1990) of oracy across subjects, and thereby capture the underlying oracy construct and norm sources.

In this research, a concrete method "(meta (after) hodos (road) (gr.))" was chosen "as to follow a path" (Nyrnes 2002, 24, my translation) into the students' landscape of knowledge where different topoi (places (gr.)/singular is topos) represent metaphorical determinations providing different perspectives (Gabrielsen and Christiansen 2010, 72). I assume that for students, schools are loci communes, common places where conforming practices exist (Tønnesson and Sivesind 2016, 202). Topoi are structures of thoughts (mind maps) that knowledge originates from and where relational concepts assist in understanding the students' schools as cultural contexts, with ideology, norms, and doxa (Togeby 2004). In order to reveal doxa, rhetoric becomes a helpful tool; "rhetoric is the means, doxa is the goal" (Gabrielsen 2008, 65, my translation). Doxa represents the students' reality, their view of the world, their doxa. A rhetorical topos analysis was performed to uncover the distribution of topoi in six focus group interviews with 22 tenth-grade students representing two lower secondary schools in Norway. Firstly, I carried out an abductive qualitative rhetorical topos analysis (abductive strategy 1) (Berge 1996, Svennevig 2001). Secondly, I searched for patterns in the quantitative distribution of the different types of topoi (abductive strategy 2). I created two research questions in order to achieve my objectives:

1) How do tenth-grade, lower secondary, Norwegian students perceive the assessment and teaching of oracy, and the underlying oracy construct across subjects as a key competence?
2) What norm sources do the students refer to in their work with oracy?

The field of Language Assessment Literacy (LAL) incorporates knowledge and competencies that stakeholders like students and teachers should possess about language assessment (Stiggins 1995); however, "language as a construct for assessment is what differentiates LAL from its generic term" (Giraldo 2018, 183). Previous research in language assessment literacy indicates that in the process of becoming lifelong learners, it is crucial that the students become involved in their own assessment situation (Stiggins 2012). The students achieve more when they comprehend what is expected of them, and when they understand their own level of performance (Engelsen and Smith 2014, Stiggins 2006). A classroom atmosphere, where the students and the teachers have a common understanding of assessment and "speak the same language in assessment", enhances the students' learning (Engelsen and Smith 2014, 93). The teachers' conceptual understanding of oracy and the communication thereof (Giraldo 2018) might therefore be essential in the students' learning processes. In order to build students' confidence and maximize students' mastery, students should be involved in their own assessment defined as assessment for learning (Stiggins and Chappuis 2006, 11). According to Stiggins and Chappuis (2006), it is crucial to teach students sufficient competencies to achieve mastery through goal-setting, self-assessment, and reflection as well as providing opportunities for students to share their learning. Language assessment researchers call for more assessment research from the students' perspectives (Gamlem and Smith 2013, Hattie and Gan 2011).

Preceding language assessment literacy investigations on the oracy construct have limitations due to the particular group of teachers being targeted i.e. language teachers and especially English as a second language teachers (L2) (Bøhn 2016, Luoma 2004, Tsagari and Vogt 2017). Since the participants in the research have mainly been language teachers, stakeholders such as students, parents and teachers in subjects other than languages, have received very little attention (Malone 2016). Students' attitudes and opinions only seem to receive limited attention in research connected to the teacher-student interaction in assessment (Leung and Mohan 2004, Kontovourki 2012). The question arises as to whether or not this existing research can be generalized to oracy as a key competence setting (Kaldahl 2019). This might especially be a concern due to the fact that each individual subject has a discipline tradition which might be better fitted for an input-output assessment tradition. This is in comparison to a cross-disciplinary competence such as oracy which is more generic and more related to the real world (Kaldahl 2019). Thus, research on the oracy construct across subjects is needed, especially investigated from the students' perspective. The research presented in this chapter addresses this gap and provides new insights to LAL.

Since how to talk well cannot be individually understood (Berge 2009, 17, Quintilian 2004), Berge (1990) has used the term 'text norms' to characterize what humans, in a certain community or culture, view as good quality utterances in assessment situations. Even though these theories are based on teachers in school cultures and assessment situations, I hypothesize that the term text norms can be utilized to explain students' oracy norms in a classroom or school community. In rhetoric, being able to express oneself in accordance with the topic and the context is closely related to the term *bildung* (paideia (gr.)) (Berge 2009) which from a rhetorical point of view represents the overarching purpose of education. The introduction of oracy as a key competence in the curriculum, becomes something more than oracy in each subject discipline, it becomes the foundation for students to develop both intellectually and emotionally; thus, reintroducing the progymnasmata or the rhetorical educational view (Berge 2009, 17). This is closely related to the term rhetorical citizenship, as understood through the ability to find arguments and utter oneself through participation in the democracy, as well as being emphatic and valuing another's point of view; but also including the participatory aspects by recognizing manipulation and persuasion (Kock and Villandsen 2014, 18). How people gain access to civic life is defined through rhetorical agency (Kock and Villandsen 2014, 10). Klafki (2014) adapts the rhetorical paideia in his *bildung* where the purpose is solidarity with other humans in pursuing societal rights through self-determination and co-determination. Klafki (2014) divides the concept of *bildung* into material *bildung*, the content as objective information of the world, and the subjective side, where knowledge becomes meaningful for the individual (categorical *bildung*). According to Klafki (2014, 120) only the latter contributes to *bildung*, in a reflective-dialogue where the world opens up for the student and the student opens up for the world.

A country's school visions are often represented through the national curriculum. Goodlad's (1979) theory of the curriculum's five faces, describes the curriculum's journey. Firstly, from the political ideology behind it. Secondly, to the formal curriculum; thirdly, to the perceived curriculum, and in the fourth face to the teachers' interpretation of the curriculum through their practical teaching. In this face we find the teachers' enactment of policies, where in return there might not be a direct link to policies (Braun, Maguire and Ball 2010). The fifth and final phase is where we find each individual student's understanding of the experienced curriculum (Goodlad 1979). In this study the context is the students' actual social and cultural reality, their kairos, which in the didactical triangle illustrates the interactions between the student, the teacher and the content (Künzli 1998). In the students' kairos they express their intentional oracy or rhetoric (Fafner 2005). Thus, the researched oracy construct is not a construct in a psychometric understanding (Kane 2006) but more a quasi-construct found in the school context seen from the perspective of the students (Borgström 2014, Author 2019).

Rhetoric offers a meta language on oracy (Kaldahl 2019). A psychological explanation of how persuasion takes place can be explained within Aristotle's three modes of persuasion: logos (i.e. subject-specific content), the ability to display character (ethos) and have emotional influence on the audience (pathos) (Aristotle 2006). A persuasive speech is adjusted to the rhetor, the topic, the audience and the context in the rhetorical situation (Bitzer 1997). Persuasion mastery involves knowing what to say, express and deliver at the right moment (kairos) as well as in a suitable fashion for the context (aptum); thus, adapting the speech in accordance with the recipients' doxa (Andersen 1995, 165).

2. Methods

2.1 The Participants and the Approach

This is a qualitative study. In order to answer the research questions, answers have been sought from the students' own perceptions, experience, and reflections, as knowledge was expressed through six focus group interviews with 22 students from two different, large, urban, lower secondary schools in Norway (Brinkman and Kvale 2015). The focus groups varied in size from two to five students with a mix of gender (14 boys and 8 girls), based on volunteerism. As a group the students represent every single subject taught at the lower secondary level.

The primary data are interviews based on a semi-structural interview guide with 12 open questions, conducted in the spring of 2016. The semi-structured interviews gave room for follow-up questions and lasted 20-60 minutes depending on how talkative the students were. The material was recorded and transcribed verbatim. In the abductive qualitative rhetorical topos analysis (abductive strategy 1), I thematically analyzed patterns in the material (cf. Bakken 2006), by asking: "What is said here regarding i.e. the assessment of oracy?" (cf. Tønnesson 2004 in Tønnesson et al. 2016, 208). Each sequence in the interview with continuous speech between two interruptions caused by another student or me (the interviewer) represents the unit of analyses. Next, patterns in the quantitative distribution of the different types of topoi (abductive strategy 2) were identified and organized into groups of topoi; thereby, developing topoi lists. Lastly, I generated a common meaning from the content of the textual units. In order to create transparency, Table 1 aims to illustrate the interpretative analysis and Figure 1 shows the associated topoi lists. Throughout the entire analytic process, findings were validated by systematically comparing categories and content.

Utterances:	Associated main categories of topoi:	Associated subcategories of topoi:
Good oral competence is to have the ability to express ones opinions and informative knowledge to others, so that other people comprehend and learn. Ones abilities to argue and deliver informative knowledge is most important.	Logos: Aspects with the content in students' oral texts Pathos: The students' abilities to awakening the audience's attention	L2= express L29=knowledge L23=argue/discuss P1= receiver awareness
It seem more trustworthy, persuasive, and it easier for the listener to follow. However, it becomes less challenging for us to express body language after years of practice and feedback on this.	Ethos: Aspects with the students' body language and voice Didactics feedback	E11= credible/trustworthy E5=body language P5= persuasive P1= receiver awareness D3= practice F7= feedback
Some teachers value how much you participate in class like how often you raise your hand.	Frequency and quality: Aspects with students willingness to participate	FQ1= participation/activity/raise your hand
It is receiving a grade that makes you nervous.	Challenges (challenges with oracy)	C3= anxiety
We learn a lot from each other's presentations.	Norms (contexts where the students learn oracy)	N3= learn from each other's presentations

Table 1: Utterances and associated topoi

TOPOI

Logos (language/content): Formulate, express (meaning), understandable, short and precise, be concrete, good at holding a speech, read, facts, talk, have an outline/plan/disposition for what to say, stick to the plan, intonation, memorize, able to have spontaneous responses, content, reflect, cause effect, inform, visual aids, present, language, add sources, argue/discuss, pronunciation, grammar, vocabulary, subject terminology, knowledge, improvise, converse (lead), communicate, interesting, explain (33)

Ethos (body/voice): into, passionate/engaged, care about the topic, stand up for what you mean, body language, posture, take the stage, sure of yourself/assertiveness, voice, talk loud and clear, be credible, free yourself from script, exciting, eye contact, use of humor, first impression, dramatize (17)

Pathos (room/time/situation): receiver awareness, exciting to listen to, proper, persuasive, say the right thing at the right time (aptum), please the teacher/be strategic (6)

FQ (frequency and quality): participation/activity/raise your hand, right answer, minutes/time, used to weight grade (4)

Challenges (challenges with oracy): dyslexia, some subjects are more challenging to be oral in since some subjects have just one answer, anxiety, time, to be audience for many presentations in a row is tiring, warm and accepting atmosphere, anxiety for grading, lack a recipe on how to perform oral texts, self-confidence, shy, writing is more in focus than oracy, okay to do well in school (12)

Norms (possible contexts for learning oracy): focus on oracy vary from teacher to teacher, L3 teacher focuses on oracy, learn from each others' presentations, various practice between subjects, social science encourages oracy, science do not encourage oracy, math do not encourage oracy, L1 encourages oracy, L2 encourages oracy, oracy in math is to talk about an assignment or solve it on the black board or answer questions, PE encourages oracy, art does not encourage oracy, criteria, better at holding presentations in 10th grade (progression), learn from teachers' feedback, learn from teachers/principal as role models, learn from parents as role models, previous experience or not from practice non practice on presentations in elementary school, religion encourages oracy, learn from practicing oracy, oracy useful in adult life (work/politics), rhetoric, learn from internet, friends, L3 no focus on oracy, patterns for oracy in all subjects, rhetoric in textbooks used for writing, born with oral competence, teachers agree on oracy, PE does not encourage oracy, food and health subjects encourage oracy, oracy in textbooks, learn when writing script, self-learned, art encourage oracy (36)

Feedback (response on work): on learning platform, various feedback from teacher to teacher, very limited information in feedback from teacher, ask specifically the teacher for the reasoning behind a given grade, ask specifically the teacher for the reasoning behind a given grade, peer assessment, feedback from teachers (assessment), immediate oral feedback, oral feedback, feedback from teacher a long time after, concrete teacher feedback, immediate feedback, individual more personal feedback, teachers agree on assessment (14)

Didactics (ways to work with oracy): presentations, debate/discussion, practice to prepare for oral exam, group work/cooperation, teacher encourage oral presentation/conversation, dialogue, rhetoric used for text, rhetoric used for oral text, more variations on themes on presentations, practice presentations in front of parents, conversations, rhetoric used for written assignments, practice in front of friends, practice presentations on your own, spontaneous (do not practice presentations), prepare power points, drama, read, ask questions, make a film (20)

Figure 1: Topoi Lists

In terms of ethical considerations, written consent forms were signed by the students and their parents, detailed characteristics of the participants were left out, the materials were handled anonymously, and the study was approved by the Norwegian Centre for Research Data.

2.2 The Norwegian Context

Norway has a population of five million, and its educational system offers students (in the age range of 6-19) thirteen years of education at no cost. The educational system is organized into primary school (grades 1-7), lower secondary school (grades 8-10) and upper secondary school (grades 11-13) (Hertzberg and Roe 2015). The Ministry of Education voted by Parliament prepares national steering documents, such as the National Curriculum and the assessment system, which serve as the framework for school owners, schools, and classroom teachers, which again affect each individual student (Hertzberg et al. 2015). The students are not subjected to grades in primary school. There are national tests in grades 5 and 8/9 in mathematics, English and reading. The students are given mid-term and end-of-year evaluation grades based on overall achievement grades in all subjects through classroom assessments throughout the academic year. In addition, the students have school exit exams in grades 10 and 13 (Hertzberg et al. 2015). Some of the subject disciplines like Norwegian as a first language (L1) and English as a second language (L2) have a separate oracy grade; additionally, all 10th grade students have to take one written, and one oral high-stake national exam at the end of the 10th grade, these exams are both based on summative assessment (Bøhn 2016). The oral exam in Norway is a performance assessment in order to find out whether the student can perform a certain task (Kane 1999). The final grades from the 10th grade, both the end-of-year achievement grades and the grades from the written and oral national exams, will determine whether the students will be admitted to the general studies program or various vocational upper secondary education programs; it is thus a competitive process for the students (Bøhn 2016). The exam evaluation is based on the curriculum context, which in return is intrinsically linked to the teaching-learning processes in the classroom (Bøhn 2016).

2.3 Methodological Limitations

One challenge with this method is that the material is gathered through interviews with the students, and not through observations of the students in the actual assessment practices. The fact that the students in their answers might present their perceptions of good assessment practices instead of the actual assessment practices might be a dilemma. However, in this study I will not go further into the distinction between the teachers' actual oracy assessment practices and the perceived ones, but present the gathered information as the students' perceptions of the teachers' assessment practices as oracy as a key competence. Since oracy is traditionally different in each subject, and there exist different understandings of the assessment of oracy even within each subject discipline; the material that I gather will be anchored in the students' tacit knowledge based on everyday practices, different school cultures, and different subject disciplines (Matre et al. 2011).

3. The results, analysis and discussion

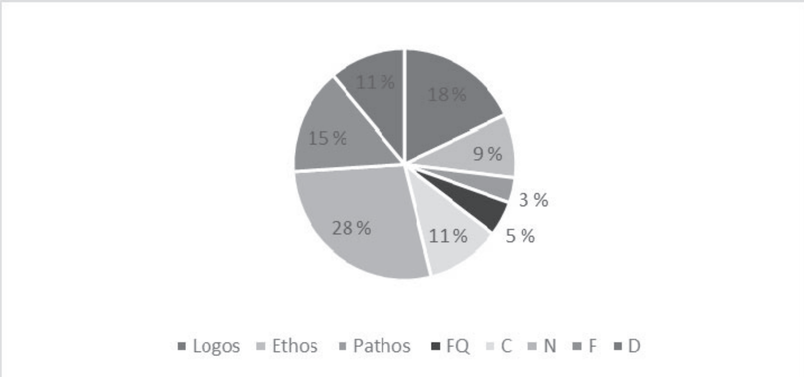


Figure 2: Students' Overall Topic Landscape of Oracy

At the level of topos groups, the most frequent of all main topoi/topos groups was norm sources (28%), illustrating where the students report they have learned about oracy; however, there are 37 topoi in the category of norm sources. At the level of a single topos, the students most often stressed various practices between subjects. The students expressed how they have to find out what good oracy looks like in every subject and even between each individual teacher representing the same subject:

> In a way it is easier with oracy in the social sciences and in religion compared to i.e. mathematics. There is more room for discussion and more than one answer; however, in mathematics there is only one right answer. I believe this limits oracy in mathematic lessons.

The students also address that some teachers, when assessing oracy, award oral participation in class measured by how often the students raise their hands and show eagerness to participate (cf. frequency and quality, 5%). The students have experienced that this frequency measure has been used when the student's achievement is between two different grades to either lift the student up to the higher grade or use this frequency measure to argue for giving the student a lower grade. The students have also experienced teachers who do not bring up participation in class as a criterion for the grade or adjust the grade up or down a level; however, these teachers seem to focus only on the prepared oral presentations the students have in class as the basis for the grade. One student shared: «Some teachers value how much you participate in class like how often you raise your hand.”

When the students were asked where they learned oral competence as a key competence (norm sources, 28%), one answered:

> Most of my good oral competence, I have learned at home. It is evident that you learn oracy when you are young in meetings with other people. We were taught oracy in kindergarten and in elementary school. Now we learn oracy especially in Norwegian as a first language (L1), and from the English teacher in English as a second language (L2). But, otherwise I would like to stress that I learn most English from the internet, I constantly keep myself updated on English oracy on the internet.

Students' norm sources represent a broad spectrum of contexts where they learn oracy or notice good rhetorical pointers. The students use the internet as a source, they view other people and mention Obama as a good rhetorician, but also the Norwegian king, who is known for his New Year's Eve speeches. However, the students claim that he could be better at freeing himself from the script probably due to his age and possible dyslexic diagnosis. Students from one of the schools tend to admire their principal for their excellent rhetorical competences, when talking to the students at the school as a whole group, since the principal has a clear and concrete message and leaves no questions unanswered. Several students also mentioned teachers as good role models of speech but also their parents as well as fellow students. Students are aware that oracy is often acquired at home in their families. Students also practise oral presentations in front of their parents and write manuscripts with their parents, which they in return stress they learn a lot from. However, one student expressed:

> I cannot practice oracy in front of my parents, since they are from Poland. They do not speak Norwegian, and I do not get much help from them with homework. I usually practice on my own or in front of a classmate.

This Polish student stresses the issue of cultural capital or habitus where an educational system that does not teach students rhetorical competence, reproduces structures of inequality in society. Young people, who do not acquire knowledge about how they should utter in accordance with the accepted norms for a context or a culture, are at risk for missing the opportunities to influence society through agency and rhetorical citizenship. In return, this accentuates rhetoric as the very core of *bildung* in education.

Another student supports this:

> If you do not come from a family where good oracy or conversations are in focus, then you do not know how to express yourself correctly in other contexts either.

The latter quote underscores an ecological approach to oracy, which indicates that an educational construct of oracy should be defined in relation to real-life contexts (Berge et al. 2016).

Students also believe that they learn a lot from working in groups and viewing other students' presentations or listening to fellow students' feedback from peers and teachers. One student said: "We learn a lot from each other's presentations." Furthermore, the students are welcoming constructive and informative feedback in a timely fashion. If it takes too long for feedback, the students claim that they forget and only care about the grade.

> If you have to wait for weeks on the feedback or maybe a month, you start to forget, and then you only care about the grade.

> Feedback has to be informative and constructive, more than just general, like just okay in a way. You need to have feedback in order to feel a sense of mastery, so you can build self-confidence and do better next time.

However, in the feedback category (cf. feedback, 15%), students value the fact that they have teachers who are extremely careful in giving feedback or critical pointers in whole class situations. In peer feedback, it is easiest to give good feedback to peers, leaving it up to the teachers to communicate things that should have been done better. Student-made decisions show evidence of ethical judgements made in connection to the context:

> We are very considerate, and do not say something which can be detrimental for other students, which consequently can make them anxious of presenting orally again.

A safe environment where the students are provided with sufficient time to develop makes it easier to dare to share oral utterances and own personal opinions or unique views of matters. The students also express that being in the same class at lower secondary school for three years makes them develop trust (pistis) with the teacher and even between the students. They view this as essential for a community of sharing and learning and the acquisition of knowledge. One student expressed:

> Students in our class stutter less (read: less anxious) when we present now in 10th grade (implicit compared to 8th grade), we are in a way more comfortable with each other.

Students do not believe that they receive education in oracy, for them oracy is more an activity without a clear theory behind it. They miss a "recipe" on how to do oral presentations correctly or in other words a meta-language on oracy is lacking. Most students have heard about rhetoric and know what it is, but they claim that the only chapter they have about rhetoric is in an L1 textbook, which in turn is used for written text analysis and making commercials. The teachers do not seem to utilize the potential of this rhetorical chapter as a meta-language for oracy. One student expresses this:

> Through three years of lower secondary, I have had a lot of oral activity in class especially as presentations, but we have never received instructions in or a throughout review of what a good oral presentation really is.

Another student supports this finding:

> The knowledge on oracy is like this: Find out how to do it on your own!

The last quote reveals that oracy as an important cognitive and sociocultural dimension of the students' learning process is not taken good care of in Norwegian schools, and the norms may not be necessarily represented as formalized genres. Thus, I found it necessary to ask the students what the teachers do value when assessing oracy.

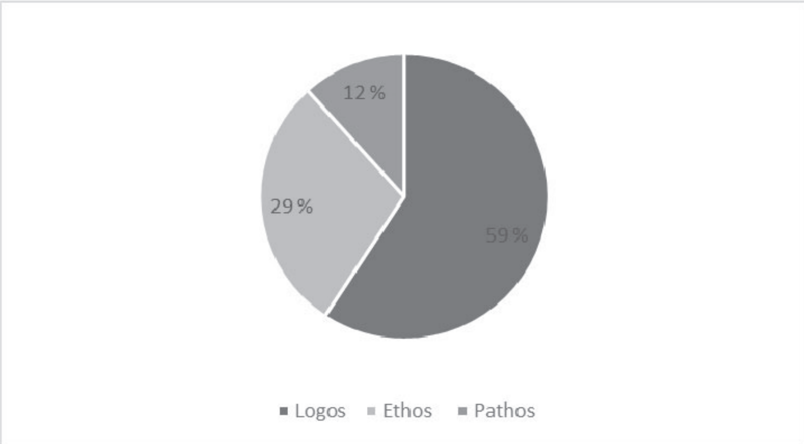


Figure 3: Students' Oracy Construct

The results indicate that the content and language aspects (cf. logos, 59%) of the students' utterances are the most valued parts of the oracy construct. The students' abilities to display character through body and voice (cf. ethos, 29%), and the students' abilities to appeal to the teachers' sense of emotions (cf. pathos, 12%) are two other important dimensions. The students acknowledge that some subjects are more oracy-oriented than others, while some teachers, even though they teach the same subject, view, emphasize and encourage oracy differently. The students see a pattern of an oracy construct across subjects:

> There are a few criteria that repeat themselves across subjects, like a pattern for oracy that several teachers appreciate.

When the students were asked to define what having good oral competence as a key competence entails, one student expressed:

> Good oral competence is to have the ability to express one's opinions and informative knowledge to others, so that other people comprehend and learn. One's abilities to argue and deliver informative knowledge are most important.

This situation illustrates that the students know that knowledge has to be displayed in an argumentative and informative matter (logos elements); yet, the student is very receiver-oriented (pathos) and is concerned with the context and opportunity for fellow students to learn. He continues:

> Oracy in school is to talk about subject knowledge, so we can learn from each other and from ourselves. In return, we are able to utter ourselves outside of the school context in the society at large.

The students view the school as the arena of an accepting and warm environment where learning, through making knowledge your own, as well as practice can take place in order for them to become future citizens with agency and rhetorical citizenship:

> We become more comfortable at expressing ourselves, and we are enabled to contribute with our own opinions. In this way, it becomes easier in the future to hold job presentations, and to be able to choose positions in the management. In a way, it becomes easier to participate. If you work in the municipal council, it becomes easier to express yourself. We exercise oracy from kindergarten, but now in 10th grade we exercise to utter and argue professionally through delivering content knowledge.

This student uncovers the very purpose for rhetoric in schools, by viewing oracy as a channel to deliver utterances of their own, personable and unique opinions to others. The school is his arena for training; however, he views the complete educational setting, even kindergarten, as a stage for practising future rhetorical citizenship. This student values good oral communication as a requirement for leadership positions within the field of work as well as in governmental positions and that the educational system plays an essential role here in preparing future generations in a democracy.

When the students were asked what the teachers are looking for when they assess oracy as a key competence, one student responded:

> It is really important with reflections. It is crucial to reflect around the presented facts, so you really show that you know a lot about the topic… Cause and effect relations are important to present. In addition to knowing the content, it is important to be good at expressing yourself. Because if you only know and present facts point by point, your presentation will not be so copious and exciting as it could have been if you were able to reflect and express yourself like you meant it.

This student also stresses the logos elements of the utterances, and views the ability to understand cause and effect as crucial in knowledge attainment. This student weights the importance of uniqueness, newness and original utterances. Another student goes on:

> Body language is very important when you present, because you are in front of the classroom and everybody is looking at you. It is crucial to show some movement, gimmick, and enthusiasm. It seems more

> trustworthy, persuasive, and it is easier for the listener to follow. However, it becomes less challenging for us to express body language after years of practice and feedback on this.

Action qualities like body language, gestures, mimics, voice and eye contact (Gelang 2008) are valued among the students. They even talk about being trustworthy and persuasive in rhetorical contexts. The use of humour and acting is appreciated as components of students' oracy:

> My language competence in German (L3) is more like a C; however, by doing some acting and using humour, I received a B plus.

Students acknowledge oracy as a cultural phenomenon, and they are receiver-oriented. Many students are very strategic and they aim to please the audience, and in the school context first and foremost the teacher. The students know that their grades and future education are at stake:

> In order to receive a good grade it is smart to know what your teacher is interested in. If she is interested in politics, you should include politics in your presentation. If your teacher is a real football fan, it is probably strategic to mention something with a football team.

They are also very aware of the receivers' age group:

> It is all about keeping the balance between sharing important facts, concretely, but also to use a sense of humour with the class, because we are 15-16 year-olds we do not care one hundred per cent about facts, but how we deliver them.

What worries the students the most in an oral assessment situation is not presenting in front of their classmates, but to have a teacher present in the classroom who is ready to assess the quality of their performances:

> What creates anxiety with oral presentations is the very fact that the teacher is sitting there, taking notes and assessing. Our classmates have been there for three years, we know them well. We realize that to speak in front of them is not viewed as a problem, they will not laugh of us.
>
> It is receiving a grade that makes you nervous.

However, oracy assessment provides the students with an opportunity to shine in an assessment situation, and to feel a sense of mastery:

> It is crucial to have fun in school. I am really looking forward to the oral exam. Maybe others are nervous, but for me it is plain fun, since I am good at oracy, and I am one of the worst students in the class in writing.

4. Conclusion and implications for further research

These students show evidence through their meta-cognitive abilities to reflect and be critical and conscious around their own learning processes and contexts for oracy. These competencies are important in achieving ultimate academic success as well as in their future personal and professional lives. This illustrates the necessity of integrating students in their own learning through formative assessment processes such as peer feedback and self-evaluation with a meta-language on assessment as well as a meta-language on oracy.

The students view the school setting as a place to practise in a safe and warm environment, where knowledge can be acquired through teachers' facilitation of sufficient time and attentiveness for each individual to develop mastery. They acknowledge that oracy has both a cognitive as well as a sociocultural side. The students naturally view oracy as a form of categorical *bildung*; however, in some assessment situations the concept of material *bildung* becomes more apparent. The very goal for oracy for students becomes democratization, the ability to value, apprise, and take other perspectives, as well as being empowered to take personal stands and utter themselves.

Meaningful work with oracy is crucial for students to avoid surface learning, acquire personable knowledge and form deeper learning. Furthermore, the students view oracy as a supportive learning tool for intellectual stimulus. For them, oracy is important within school as well as outside school, as empowerment through agency and rhetorical citizenship. They know that good oral competence is key to the field of work and leading political positions, communication in families as well as in professional settings and for democracy. According to the students, oracy becomes something more than just oracy for assessment in the subject. It becomes the very core of *bildung*.

The rhetorical vocabulary has been valuable in this work; however, a meta-language on oracy is lacking in Norwegian schools. This raises several issues. How can the quality of students' utterances be assessed in a credible and learning supportive way? Do students receive sufficient information and knowledge from the teachers to be able to utter themselves with high quality in the subjects? Does the current assessment of oracy withstand a psychometric control? Systematic training in oracy is needed. Rhetoric offers a theory which could serve as the basis for curriculum development, textbooks, teaching and be the "the recipe" for good rhetoric that the students request. The students do not receive sufficient training or knowledge in rhetorical competence; consequently, societal inequalities will be reproduced. The ones from homes without the "right" rhetorical competence or the gift of speech will lose out on the possibilities to influence society and become targets for manipulation and persuasion. This can be detrimental for the new generations who are constantly subjected to amounts of new information though media and the internet. It is currently, impossible to know what knowledge the students who go through the educational system today will need in the future; on the contrary, the students know that good oracy or good rhetorical competence is the very stepping-stone into the future with agency, rhetorical citizenship and possibilities for a strengthened democracy.

The students show evidence of intrinsic motivation; they see the purpose and usefulness of oracy. However, the assessment of oracy situations seems to reduce the complexity of real-life situations and can be seen as insufficient and deductive to learning. The grading or the judgement is often conceptualized by the students as valuing the product and not the process as well as subjecting the students to anxiety in test situations. The students are process-oriented and learn democracy in the process. The goal- and result-oriented assessment creates anxiety and seems to reduce students' intrinsic motivation and might be detrimental to students' inductive learning processes.

For the students, the process of learning oracy is universal and oral competence is for the students highly valued as a generic competence. The content in each individual subject is dealt with by the students in a rational and pragmatic matter. The students make their own reflections, logical rationales, choices and values inherent in the learning process. These student-made decisions also include ethical judgements made in the clarification of the particular context, which is an advanced form of knowledge and represents deep learning.

From the students' point of view, oracy is one of the main means for the acquisition of knowledge and judgement, and is inherent in the process as concrete and practical exercises. Due to different subject traditions, students know that assessment varies between disciplines and that this cross-disciplinary competence is more related to the real world. It becomes almost contradictory to assess a real-life competence with an input-output measure.

Additionally, it seems as if these students become resourceful when exposed to oracy exercises with no clear recipe. When the school does not provide a meta-language for oracy, students tend to use their common knowledge or experience-based knowledge, independent of school. Hence, it can be interpreted that teachers do not communicate a conceptual understanding of oracy and the assessment thereof (Giraldo 2018).

Consequently, a conceptual understanding of oracy and its assessment needs to be developed in the area of teachers' professional development.

It is almost contradictory, but it seems as if the lack of teaching becomes a resource for the students' learning. However, this raises questions according to how students, who do not have the same experience-based knowledge and/or have fewer resources, solve being challenged with these tasks. In sum, by being exposed to teaching and feedback without a meta-language on oracy, and consequently leaning on students' common knowledge, social and relational aspects of knowledge are amplified, thus risking reproducing social inequalities.

5. References

Andersen, Øyvind. 1995. *I retorikkens hage* [In the garden of rhetoric]. Oslo: Universitetsforlaget.

Aristotle. 2006. *Retorikken* [Rhetoric]. Oslo: Vidarforlaget.

Bakken, Jonas. 2006. «Litteraturvitenskapens retorikk. En studie av tekstnormene for gode norske empiriske litteraturvitenskaplige artikler i perioden 1937-57» [The Rhetoric of Literature. A Study of the Text Norms for Good Norwegian Empirical Literature Articles in the Period 1937-57] (PhD Thesis). Oslo: University of Oslo.

Berge, Kjell Lars. 1990. *Tekstnormers diakroni: noen idéer til en sosiotekstologisk teori om tekstnormendring* [Diachronia of Text Norms: some Ideas for a Socio-textual Theory of Textual Change]. Stockholm: Stockholm universitet, Institutionen för nordiska språk.

—. 1996. «Norsksensorenes tekstnormer og doxa. En kultursemiotisk og sosiotekstologisk analyse» [Norwegian Sensors' Text Standards and Doxa. A Cultural Semiotic and Socio-textological Analysis] (PhD Thesis). Trondheim: Institutt for anvendt språkvitenskap, Det historisk-filosofiske fakultet, NTNU.

—. 2009. "Writing as a Text-cultural Competence: Challenges and Solutions in Defining Writing as a Basic Competence in the New Norwegian Curriculum." In *The Diversity of Speech and Writing*, edited by J. Kalliokoski, T. Nikko, S. Pyhaniemi, and S. Shore, 15-42. Helsinki: AfinLa.

Berge, Kjell Lars, Lars Sigfred Evensen, and Ragnar Thygesen. 2016. "The Wheel of Writing: a Model of the writing domain for the Teaching and Assessing of Writing as a Key Competency." *The Curriculum Journal* 27(2):172-189.

Berge, Kjell Lars, Gustaf Bernhard Skar, Synnøve Matre, Randi Solheim, Lars Sigfred Evensen, Hildegunn Otnes, and Ragnar Thygesen. 2017.

«Introducing Teachers to New Semiotic Tools for Writing Instruction and Writing Assessment: Consequences for Students' Writing Proficiency." *Assessment in Education: Principles, Policy & Practice*: 1-20.

Bitzer, Lloyd Frank. 1997. "The Rhetorical Situation." *Rhetorica Scandinavica* 1-14. 1997/3:9-17.

Borgström, Eric Hans Gustaf. 2014. «Vad räknas som belägg för skrivförmåga? Ett textkulturellt perspektiv på skrivuppgifter i den svenska gymnasieskolans nationella prov» [What is Written Competency? A Text Cultural Perspective on Writing Assignments on the Swedish Upper Secondary National Test]. *Sakprosa* 6(1):1-34.

Braun, Anette, Meg Maguire, and Stephen J. Ball. 2010. «Policy Enactments in the UK Secondary School: Examining Policy, Practice and School Positioning." *Journal of Education Policy* 25(4):547-560.

Brinkmann, Svend, and Steinar Kvale. 2015. *Interviews: Learning the Craft of Qualitative Research Interviewing*. Los Angeles: Sage.

Bøhn, Henrik. 2016. "What is to be Assessed? Teachers' Understanding of Constructs in an Oral English Examination in Norway" (PhD Thesis). Oslo: University of Oslo.

Creswell, John W. 2014. *A Concise Introduction to Mixed Methods Research.* Los Angeles, CA: Sage Publications.

Engelsen, Knut Steinar, and Kari Smith. 2014. «Assessment Literacy." In *Designing Assessment for Quality Learning*, edited by Claire Wyatt-Smith, Valentina Klenowski, and Peta Colbert, 91-107. Netherlands: Springer.

Fafner, Jørgen. 2005. *Retorik: klassisk og moderne: indføring i nogle grundbegreber* [Rhetoric; Classic and Modern: Introduction to some Basis Concepts]. København: Akademisk Forlag.

Gabrielsen, Jonas. 2008. *Topik. Ekskursjoner i den retoriske toposlære* [Topik. Excursions in the Rhetorical Topos]. Åstorp: Retorikforlaget.

Gabrielsen, Jonas, and Tanja Juul Christiansen. 2010. *Talens magt* [The Power of Speech]. Hans Reitzels Forlag.

Gamlem, Siv M., and Kari Smith. 2013. «Student Perceptions of Classroom Feedback." *Assessment in Education: Principles, Policy & Practice* 20(2):150-169.

Gelang, Maria. 2008. "Actiokapitalet–retorikens ickeverbala resurser" [Actio-capital–the Nonverbal Rhetorical Resources] (PhD Thesis). Örebro Universitet, Åstorp: Retorikförlaget.

Giraldo, F. 2018. «Language Assessment Literacy: Implications for Language Teachers.» *Teacher Professional Development* 20(1):179-194.

Goodlad, John I. 1979. *Curriculum Inquiry*. New York: McGraw-Hill Book Company.

Hattie, John, and Mark Gan. 2011. "Instruction Based on Feedback." In *Handbook of Research on Learning and Instruction*, edited by Richard E. Mayer, and Patricia A. Alexander, 249-271.

Hertzberg, Frøydis, and Astrid Roe. 2016. "Writing in the Content Areas: a Norwegian Case Study." *Reading and Writing* 29(3):555-576.

Kaldahl, Anne-Grete. 2019. "Assessing oracy: Chasing the teachers' unspoken oracy construct across disciplines in the landscape between policy and freedom." *L1-Educational Studies in Language and Literature* 19:1-24. (Special issue on Assessing Oracy, edited by Anne-Grete Kaldahl, Antonia Bachinger, and Gert Rijlaarsdam.)

Kane, Michael, Terence Crooks, and Allan Cohen. 1999. "Validating Measures of Performance." *Educational Measurement: Issues Practice* 18(2):5-17.

Kane, Michael. 2006. "Validation." In *Educational Measurement*, edited by Roger L. Brennan, 4(2): 17-64. Westport: CT: Praeger.

Klafki, Wolfgang. 2014. *Dannelsesteori og didaktik–nye studier* [Formation Theory and Didactics–New Studies] (3rd ed.). Århus: Klim.

Kock, Christian Erik J., and Lisa Storm Villadsen. 2014. "Rhetorical Citizenship as a Conceptual Frame." In *Contemporary Rhetorical Citizenship*, edited by Christian Erik J. Kock, and Lisa Strom Villandsen. Leiden University Press.

Kontovourki, Stavroula. 2012. "Reading Leveled Books in Assessment Saturated Classrooms: A Close Examination of Unmarked Processes of Assessment." *Reading Research Quarterly* 47(2):153-171.

Künzli, Rudolf. 1998. "The Common Frame and the Places of Didaktik." In *Didaktik and/or Curriculum: An International Dialogue*, edited by Bjørg Gundem, and Stefan Hopmann, 29-45. New York: Peter Lang.

Leung, Constant, and Bernhard Mohan. 2004. "Teacher Formative Assessment and Talk in Classroom Contexts: Assessment as Discourse and Assessment of Discourse." *Language Testing* 21(3):335-359.

Malone, Margaret E. 2016. "Training in Language Assessment." In *Language Testing and Assessment, Encyclopedia of Language and Education*, edited by Elena Shohamy, Lair May Or, and Stephen May, 225-239. Springer International Publishing.

Marton, Ference, and Roger Säljö. 1976. "On Qualitative Differences in Learning: I–Outcome and Process." *British Journal of Educational Psychology* 46(1):4-11.

Matre, Synnøve, Kjell Lars Berge, Lars Sigfred Evensen, Rolf Fasting, Randi Solheim, and Ragnar Thygesen. 2011. «Developing National

Standards for the Teaching and Assessment of Writing." Rapport frå forprosjekt Utdanning 2020.
Mercer, Neil, Paul Warwick, and Ayesha Ahmed, 2017. "An Oracy Assessment Toolkit: Linking Research and Development in the Assessment of Students' Spoken Language Skills at Age 11-12." *Learning and Instruction* 48:51-60.
Norwegian Knowledge Promotion. 2007. Oslo: Utdannings Direktoratet.
Nyrnes, Aslaug. 2002. «Det didaktiske rommet. Didaktisk topologi i Ludvig Holbergs Moralske tanker» [The Didactic Room. Didactic Topology in Ludvig Holberg's Moral Thoughts] (PhD Thesis). Bergen: Universitetet i Bergen.
Quintilianus, Marcus Fabius. 2004. *Institutio Oratoria. Opplæring av talaren I–III.* Translated by Hermund Slaattelid. Oslo: Det Norske Samlaget.
Stiggins, Rick. 2006. "Assessment for Learning: A Key to Motivation and Achievement." *Edge: The Latest Information for the Education Practitioner* 2(2):1-19. Bloomington, IND: Phi Delta Kappan.
Stiggins, R. (1995). Assessment Literacy for the 21st Century. *Phi Delta Kappan* 77(3): 238-245.
Stiggins, Rick, and Jan Chappuis. 2006. "What a Difference a Word Makes." *Journal of Staff Development* 27(1):10-14.
Svennevig, Jan. 2001. "Abduction as a Methodological Approach to the Study of Spoken Interaction." *Norskrift* 103:1-22.
Togeby, Ole. 2009. *Steder i bevidsthedens landskab-grene på ideernes træ* [Places in the Conscious on Landscape Branches of the Tree of Ideas]. Århus: Århus universitetsforlag.
Tsagari, Dina, and Karin Vogt. 2017. "Assessment Literacy of Foreign Language Teachers around Europe: Research, Challenges and Future Prospects." *Papers in Language Testing and Assessment* 6(1):41-63.
Tønnesson, Johan Lauritz, and Kirsten Sivesind. 2016. "The Rhetoric of the Norwegian Constitution Day: A Topos Analysis of Young Norwegian Students' May 17 Speeches, 2011 and 2012." *Scandinavian Journal of Educational Research* 60(2):201-218.
Vygotsky, Lev S. 1978. *Mind in Society: The Development of Higher Psychological Processes.* Cambridge, MA: Harvard University Press.
Wenger, Etienne, Richard McDermott, and William M. Snyder. 2002. *Cultivating Communities of Practice: A Guide to Managing Knowledge*. Boston, MA: Harvard Business School Press.
Wilkinson, Andrew. 1965. *Spoken English*. Edgbaston, Birmingham: University of Birmingham.

CHAPTER TEN

SOCIOPRAGMATIC ASSESSMENT LITERACY IN THE RUSSIAN LANGUAGE CLASSROOM

ANNA DREHER AND WOLFGANG STADLER

Sociopragmatic competence is rarely fostered and even less frequently assessed in the Russian as a foreign language (RFL) classroom, because for many RFL teachers it seems easier to check students' grammar or vocabulary than to examine sociopragmatic skills although they might see the necessity of doing so. So, the purpose of this chapter is to take a look at the awareness of RFL teachers for the pragmalinguistic and sociolinguistic components of students' language ability.

The aim is first to present the research literature on sociopragmatics in connection with Russian language assessment literacy (RLAL); next, to define the sociopragmatic construct according to the *CEFR*. We will, thirdly, examine RFL coursebooks as to whether they implicitly or explicitly contain sociopragmatic content. Fourth, we will analyze test collections under the same aspect, as we know RLAL is mainly influenced by the textbooks the teachers use and by the materials supplied by TORFL (Testing of Russian as a Foreign Language). Fifth, the test formats designed for L2 languages as presented in the research literature will be given a closer look to find out whether they need to be modified in order to serve as suitable formats for RFL tasks. Finally, our findings will reveal implications for the various stakeholders and provide them with useful information on how to cooperate to improve teaching and assess competence or refine textbooks.

1. Introduction

In a survey (Drackert and Stadler 2017, 2018) carried out among RFL teachers (N=198) in Germany, Austria, Switzerland and the province of South Tyrol (Italy), the authors conclude that RFL teachers are aware of communicative competence being the state-of-the-art goal in teaching, as is also recommended by the *Common European Framework of Reference* (*CEFR*). Pragmalinguistic and sociolinguistic competences, however, do not seem to be part of this objective. They are, if at all, part of explicit teaching, and mechanical exercises often outnumber the tasks which require a focus on meaning but not on form (cf. Comer 2012, 150).

A state-of-the-art RFL teacher would need to emphasize pragmalinguistic and sociolinguistic skills and promote cross-cultural awareness, yet the teachers we interviewed (*N*=8; time=5.6 hours)[1] argue that because of the linguistic challenges of the Russian language, learners need form-focused mechanical practice and cannot simultaneously be exposed to sociopragmatic peculiarities. This is the reason why recommendations of the *CEFR* as to how learners need to be prepared to recognize a "range of greetings" or "address forms", how to evaluate them sociologically, or how to use "registers" (Council of Europe 2001, 122) are largely ignored in the RFL classroom (cf. Stadler 2018, 312).

Teaching and testing, however, need to be "a multifaceted combination of linguistic, pragmatic, and cultural resources" (Davison and Leung 2009, 406; cf. also Inbar-Lourie 2013, 306). Researchers agree that these resources are part of (second) language competence models (cf. Hymes 1972, Widdowson 1983; Davison and Leung 2009, referring to Bachman 1990 and Canale and Swain 1981; Bachman and Palmer 1996). So, if we want RFL teachers to be assessment literate, they should, on the one hand, acquire basic knowledge about competence models and the principles of language assessment before coming to understand the skills they need to develop, and, on the other hand, build up the awareness needed to grasp and interpret the processes triggered in classrooms by the teachers' (diagnostic) feedback, by fixed-point, continuous, formative or summative assessment procedures (cf. Drackert et al., in this volume; Xu and Brown's TALiP Framework, 2016).

If, due to the current migrant situation in European countries, language teaching and testing need to shift from a monolingual communicative perspective to a multi- or plurilingual intercultural or transcultural perspective, "the context of situation and culture" (Scarino 2017, 22) will also have to play an important role in the foreign language (FL) classroom. Cultural factors are of great importance for expanding learners' sociocultural background. This is what Kramsch (2011) calls "symbolic competence", implying that speakers do not only exchange words or speech acts but also "life histories, experiences, knowledge, memories and values", which presupposes "the ability to think critically or analytically on the symbolic systems we use to make meaning" (ibid., 365).

---

[1] Currently a student at Innsbruck University is working on her diploma thesis in which she analyzes these eight interviews of RFL teachers regarding their beliefs in and practices of achievement and performance tests.

In order to combine the (socio)intercultural, (socio)linguistic, discourse and strategic competences (Canale and Swain 1981) or the pragmatic (functional, sociolinguistic) and organizational knowledge (Bachman and Palmer 1996) we will use the term "sociopragmatic" competence, which allows us to include "symbolic competence" (Kramsch 2011) to foster RFL teachers' understanding for diverse linguistic and cultural systems, which exist, for example, between Austrian learners and native speakers of the target language (TL), Russian. It is vital to teach sociopragmatics right from the beginning, since general language ability often suffers from a low level of sociopragmatic development. Learners might assume that they are able to express their needs, but at the same time might not be understood by native speakers (cf. Erofeeva et al. 2017, 104).

According to the above-mentioned interviews conducted among the Austrian RFL teachers (K1–K8), their current understanding of their jobs is shaped by a mixture of job experience and subjective theories (cf. also Drackert et al., in this volume). Teachers are also influenced by the requirements of the school type (vocational vs. academically oriented secondary schools), by legal documents, the school-leaving exam (which in Austria is not standardized in Russian as it is in English and the Romance languages), the course- or textbooks they use, by the opinions of their colleagues, and the so-called authorities and experts who are responsible for in-service training programmes. A quotation taken from Interview 3 (36:26–36:42) captures the typical opinion of an RFL teacher (K3). When asked whether she also tests sociopragmatic competence, she disagrees, although she thinks "she should because that is what is exciting in reality".

2. Sociopragmatics: Definition and Research Background

Sociopragmatic competence is indeed part of several language competence models. Hymes (1972) introduced a sociolinguistic perspective to Chomsky's competence and performance model by adding the capability of using grammatical competence in a variety of communicative situations, and Widdowson (1983) was the first to recognize the "sociolinguistic conventions" of language use. The knowledge of rules and conventions was added by Canale and Swain (1981); Bachman and Palmer (1996) later called this the capacity for appropriate use of knowledge in a contextual communicative language situation.

Timpe Laughlin, Wein and Schmidgall (2015, 6) define sociopragmatics as "the sociological interface of pragmatics", stating "a sociopragmatically competent language user [is] aware of sociocultural variables such as social distance, relative power, and degree of imposition". They are aware of the rules and conventions of "situationally, culturally, and socially appropriate and acceptable language use" (ibid.), and can resort to their own linguistic resources and miscellaneous strategies.

Ovčinnikova and Kolotova (2017, 4) consider pragmalinguistic and sociopragmatic competences as inextricably linked, suggesting that learners need to define and express the meaning of a word correctly and evaluate the situation and norms of behaviour if they are to act communicatively and avoid misunderstandings. Also, Černaja and Ivanova (2012) regard sociolinguistic and pragmatic elements as inseparable, stating that tolerance, empathy and "socio-cultural awareness" make up sociopragmatic competence (ibid., 76-7).

While there is literature for English as a second language (e.g. Timpe 2012, Kanik 2013, Roever 2012, 2014), there is hardly any subject-specific research dealing with the methodological development of sociopragmatic competence and its assessment in the RFL classroom, and especially focusing on the situation and conditions in German-speaking countries. Stadler (2015, 2018) has presented studies on the development of pragmalinguistic competence and on testing sociopragmatic skills in the RFL auditorium. He explores, for example, various forms of authenticity that play a role in teacher-learner and learner-learner interactions when choosing the appropriate material, deciding how to work with it and when while creating test tasks. Pre-service teachers' awareness of sociopragmatics is raised through the comparison of authentic, simplified or adapted texts, differing, for example, in their use or number of interjections and particles (cf. Stadler 2015). In simplified or adapted texts, these parts of speech are often deleted. Nevertheless, they are characteristic of the intonation or the tone (key) of a spoken utterance and can also be reproduced by non-native speakers (NNS). However, as his study shows, the voice authenticity of native speakers (NS) differs remarkably from that of NNS. The speech act test items analyzed by Stadler (2018) are based on the format of a Written Discourse Completion Test (W-DCT) and, as the results show, favour NS and heritage speakers (HS), while Russian learners often have difficulty in solving W-DCT items where lexical, illocutionary, and sociopragmatic knowledge is required all in one (Stadler 2015, 368).

Ovčinnikova and Kolotova (2017) turn away from dealing exclusively with speech acts, which are well investigated in Russian pragmalinguistics, if we think of *greetings, apologies, requests, complaints, offering services to* or *thanking someone*. The authors analyze discourse in video sequences or text passages by presenting the list of questions formulated by Sternin (2013) in his “scheme of complex analysis of communicative situations”, thus using less investigated pragmatic means, such as routine formulae and implicature (cf. also Roever 2011, 466), and also considering the relationship between interlocutors, their tone of voice and nonverbal elements of communication. Unfortunately, no practical example of discourse analysis is offered by Ovčinnikova and Kolotova.

Krulatz (2014a) analyzes Russian e-mail discourse and focuses on inquiries from students to professors, pointing out the differences in the strategies used by NS and NNS and identifying the sociocultural deficits even among advanced learners. In a follow-up study (Krulatz 2015), the impact of these e-mail requests on Russian NS is analyzed in terms of comprehensibility, appropriateness and politeness. It is revealed that NS have significantly higher scores than NNS as regards these aspects. The main differences in the linguistic realization of requests between NS and NNS are found in the forms of addressing the reader and in the endings of the e-mails, in presenting the reason for the request, the use of negated forms, conditionals and the case sensitivity of the personal pronoun *Vy/vy* (You/you).

Afonin (2013) sees the sociolinguistic and pragmatic components and the ability to use them as an indispensable part of intercultural communication. In his paper, he takes a closer look at the respective contents (or construct) and how to convey them in a communicative and task-based RFL classroom. Afonin offers concrete suggestions of how to develop the sociopragmatic competence in RFL classes in German-speaking countries, emphasizing that the teaching of sociopragmatic content is best achieved through exploratory learning or in cross-linguistic learning contexts, as such topics increase both the motivation and the creativity of the learners and also promote their language awareness (even in their L1). The author argues that the introduction of sociolinguistic features in the RFL classroom has also been supported by the pre-service teachers he interviewed who said that intercultural divergences in the field of sociolinguistics should be a special goal in RFL lessons, because “one does not talk to matryoshka dolls, but to real human beings” (ibid., 92). Afonin does not go into the detail of whether (and how) sociopragmatic competence should be tested or assessed, yet the reader understands that what is taught also needs checking and examining.

If we define sociopragmatics as incorporating both pragmalinguistic and sociolinguistic characteristics of a language, it is necessary to have a clear understanding of what constitutes the sociopragmatic construct.

3. The Construct of Sociopragmatic Competence

The *CEFR* distinguishes between sociolinguistic and pragmatic competence–a distinction Grundy (2014, 210) calls "an eccentric position". Kusevska (2014, 99) seems to agree when arguing that it is difficult to separate the two areas as they "necessarily overlap". If a third term, i.e. "sociocultural competence", is introduced, both teachers and learners may rightly become confused. Sociocultural competence is defined as "specific features of a society and its culture which are manifest in the communicative behaviour of the members of this society" (ibid.). So, it is important in the RFL classroom to develop sociopragmatic knowledge as a component of communicative competence and to shape and sharpen the skills, abilities and strategies learners have acquired in their first language, and which they try to transfer to other languages (cf. Kasper and Rose 2001). Unfortunately, politeness conventions, discourse features or interaction schemes (Council of Europe 2001, 130) are seldom taught. This is regrettable, as for some time now the *CEFR* has provided the basis for the Austrian Curriculum of FL in Secondary Levels I and II (cf. Gesamte Rechtsvorschrift für Lehrpläne) and is also the foundation for German language curricula. In the syllabus for first and second FLs in academically oriented secondary schools, explicit reference is made to the pragmalinguistic and sociolinguistic competence of learners in the field of the "Individual Learner's Progress", stating that apart from grammatical accuracy, also the propositional precision of utterances, the appropriate use of language and a diversity of linguistic means need to be accounted for when assessing learners' overall performance (ibid.).

If we follow Krulatz' suggestions (2014b) for teaching pragmatics, we need to raise language learners' sociopragmatic awareness, provide them with a number of TL devices and practices, and expand their perception of the TL community. This is fostered by authentic materials and considerable input before analyzing samples of language and producing output (ibid., 19).

But what exactly do RFL learners need, for example, at the B1 level? Since teachers mainly rely on the contents of the textbook they use in class (cf. Tschirner et al. 2000), and not so much on curricula, syllabi, or even the *CEFR*, we will present the sociopragmatic construct as recommended in the scales of the *Framework* at levels A1–B1 (cf. Table 1 in the Appendix), which are reached by most learners of Russian in Austrian and German schools.

One of the few Russian contributions to the development of sociopragmatic competence in the RFL classroom comes from Kolotova and Kofanova (2012), who include a "problem-oriented" task as a means of activating students' action-oriented learning. It is their aim to demonstrate in short dialogues at B1 level how requests need to be formulated differently, depending on whether learners talk to a stranger or a friend.

In Section 4, there follows a close analysis of selected RFL coursebooks to show whether and in which way the construct as displayed in the *CEFR* (2001, 2018) is targeted in these books and if "problem-oriented tasks" are included to develop (socio)pragmatic competence according to Kolotova and Kofanova's model.

4. Sociopragmatic Content in RFL Coursebooks

Three coursebooks for levels A2 and B1 were chosen to demonstrate which parts of the construct appear in these books, and which exercises and tasks convey sociopragmatic competence. Two of the coursebooks (*Otlično* [A2] and *Otlično* [B1]) are published in Germany, but are also widely used in Austria. The third coursebook (*Živem i učimsja v Rossii* [B1]) written by Russian authors prepares students for an internship in Russia. It is interesting to note that none of the textbooks uses the term "sociopragmatics", despite highlighting communicative competence and oral proficiency as the main teaching goals. The reasons remain unclear; the textbook authors seem to prefer terms like *etiquette, interaction, oral expression* and *regional* or *cultural studies* instead.

In *Otlično* [A2], the authors attach great importance to appropriate phrases, and they tap into varying construct components at A2 level: *to express one's own opinion, agreement or disagreement, thanks, sympathy, congratulations, to utter requests or complaints, to give advice, make suggestions and accept or reject them,* etc. These speech acts are rehearsed in different situations by offering basic expressions and familiar phrases, and by targeting numerous social settings. Frequent connectors and common particles are introduced and practised. Idiomatic phrases, however, are used only sporadically, as "good familiarity with idiomatic expressions and colloquialisms" is a characteristic of a proficient user's level (cf. Council of Europe 2001, 27). The authors emphasize that when assessing dialogues or role play activities, it is necessary to pay attention to the learners' monitoring of their reaction(s). In the teachers' handbook that comes with the latest version of *Otlično* [A1] (Hamann 2018, 4), the authors point out that communicative competence comprises not only lexis, grammar, pronunciation, style and register, but also cultural knowledge and interactional strategies. In their opinion, all skills and competences need to be taught from the very beginning, because "of what use is an excellent pronunciation, if you miss out on the basic rules of politeness at the appropriate moment?" (ibid.).

Otlično [B1] and its teachers' manual also refrain from using the term "sociopragmatics", yet the book offers a large number of exercises and tasks teaching sociopragmatic skills. Again, the term *regional (cultural) studies* is used to describe teaching aims such as "giving advice on learners' verbal conduct in numerous situations and encouraging them to reflect on cultural similarities" (Hamann and Augustin 2014, 3). *Otlično* [B1] also taps into the construct by practising differing language functions, by paying attention to acts of *persuading* or *convincing* the interlocutor and by *arguing for or against* something. The book deals with conversational topics and taboos, job ethics and the Russians' inclination for being late, yet only marginally covers Russian "customs and traditions". Some gestures are explained, and a few common idioms and familiar sayings are suggested for conversational use at B1 level. All in all, *Otlično* [B1] is a very communicative book due to the frequent use of role plays and open task formats.

The coursebook *Živem i učimsja v Rossii* [B1] targets students who have gone to Russia to study and live there. This is why the authors have included etiquette[2] in every single unit. Among the coursebook's teaching objectives, the authors list the development of monologues and dialogues, concentrating on exercises and tasks aimed at expanding language skills in general, and speaking competence in particular. Linguistic means with regard to selected speech acts (e.g. *expressing gratitude, wishes and congratulations, proposing a toast, how to pity or comfort someone*) are constantly expanded and refined. A particular focus is laid on how to express requests and how this is done by elderly people compared to younger speakers. Once again, the word "sociopragmatics" is not used, neither in the preface, nor in the textbook's individual chapters.

A striking contrast between both B1 textbooks lies in the register differences (neutral, formal, informal) which can be observed in the Russian coursebook *Živem i učimsja v Rossii*, while the German textbooks *Otlično* focus entirely on a neutral style, as recommended by the *CEFR*: "In early learning (say up to level B1), a relatively neutral register is appropriate, unless there are compelling reasons otherwise" (Council of Europe 2001, 120). These "compelling reasons" might be seen by the

---

[2] Forms of verbal etiquette play an important role in the Russian sociopragmatic research literature. Černaja and Ivanova (2012, 75), for example, emphasize the special importance of etiquette as one of the most distinct cultural phenomena in teaching RFL.

Russian authors in the learners' exposure to different text types or in their dealings with native speakers in the various domains of social existence, which is more likely to be the case when they live in the TL community. A further contrast is to be found in the use of proverbs, idioms or familiar quotations, which are explained in Russian, and learners are asked to think of settings and scenes in which they might want to use these phraseological units. By concentrating not only on the reception but also on the production of idiomatic expressions, the textbook at this stage clearly taps into the construct of the *CEFR's* next level of competence (B2).

5. Sociopragmatic Tasks in Coursebooks

It is remarkable that all textbooks contain a spectrum of the task formats that are discussed in the research literature and considered to be useful for the testing and assessing of sociopragmatic competences (cf. Blum-Kulka 1982, Kanik 2013, Roever 2012, 2014, for example). There are 63 such exercises and tasks in *Otlično* [A2], 37 in *Otlično* [B1], and 42 in *Živem i učimsja v Rossii* [B1], showing a clear sociopragmatic focus in alignment with the *CEFR* scales. The task formats used are Matching Tasks (MT), Sequencing Tasks (ST), (Oral/Written) Discourse Completion Tasks (Oral/Written DCT) with or without rejoinder and Role Plays (RP) in both *Otlično* [A2] and *Otlično* [B1].[3] *Živem i učimsja v Rossii* [B1] mainly uses DCT, Reverse DCT (R-DCT) and interviews. DCTs are either offered with or without a list of possible speech acts to choose from. The open task formats, guided or unguided, try to fulfil the purpose of exercising a variety of speech acts with a focus on either cultural similarities or differences. At B1 level, role plays present concrete and linear argumentative content, making the interlocutors persuade or convince each other. Some of the tasks in *Živem i učimsja v Rossii* [B1] have a merely comprehensive function: learners are asked to read miscellaneous forms of address and explain why this form is used within a certain setting (Kapitonova et al. 2006, 55).

The following example is taken from the *Otlično* [A2] workbook (Hamann 2012, 46). We will show how this matching task is set up, how the rubrics are formulated and whether the items seem internally consistent.

[3] Drackert (2018, 13-14) distinguishes between "embedded" (e.g. role plays) and "discrete" testing (DCT, MC, ST) of sociopragmatic competences.

Example 1: Matching Task

Interaktion: Was passt zusammen? Verbinden Sie (je 1 Punkt):
Interaction: What goes together? Connect (1 point each):

А	Что вы берёте на второе?	**1**	На втором этаже, слева от лифта.
Б	Вы недовольны своим номером?	**2**	Думаю, что нет. У меня очень много дел.
В	Скажите, пожалуйста, где буфет?	**3**	Спасибо, я обязательно позвоню.
Г	Во сколько вы обычно встаёте?	**4**	Нет, вдвоём. Сын уже не живёт с нами.
Д	Где вы раньше жили?	**5**	Я хочу, чтобы были семья, друзья, работа и хорошая квартира.
Е	Давай лучше пойдём в кино.	**6**	Нет, нет, всё в порядке. Только с сейфом есть проблема.
Ж	Ты сегодня прочитаешь эту книгу?	**7**	Это прекрасная идея. Давай.
З	Если будут проблемы, позвоните.	**8**	Я всегда встаю очень рано, в 5.30.
И	Что вы хотите? Что для вас важно?	**9**	Пельмени со сметаной.
К	Вы живёте в квартире втроём?	**10**	У нас была однокомнатная квартира в центре.[4]

At A2 level the matching of adjacency pairs is based on familiar situations (*eating out, leisure, hobbies, living conditions,* etc.). The rubrics are formulated carelessly, and there are no distractors to be found among the answers; the items' vocabulary stems from the personal domain, as is preferable at A2. A slight discrepancy can be noted among the utterances in the column on the left, since not only questions but also suggestions (е) and conditionals (з) are offered; therefore, the items may not be entirely reliable. Another flaw seems to be that in answer (7) the word *давай* (*let's*) in the speech act is repeated in (е), (so is the word *позвонить* (*to call*) in (з) and (3)), which might be an unintentional hint for learners.

[4] Although this task is presented in the textbook as a test task, the number of possible replies listed in the right column is the same as the number of items in the left one. However, test developers and test researchers recommend that in such task formats the number of possible solutions should be higher, i.e. one or two distractors need to be added (cf. Balychina 2006, 81).

In our survey 38% of RFL teachers (*N*=171) said they did not receive any training in testing or assessment during their pre-service education (cf. Drackert and Stadler 2017, 239) and they copy tasks from coursebooks other than the ones they use in class, duplicate test tasks from test collections or take them from the Internet (cf. ibid., 247). We may therefore rightly assume that the teachers' RLAL is formed not so much by theory and research but rather by common practice and the circumstances dictated by the profession. So, if "interaction" is used as a synonym for matching "adjacency pairs", "short exchanges" or "exchange of limited information on familiar or routine matters" without mentioning the (socio)pragmatic impact, it is not surprising that teachers believe they do not test "sociopragmatic competence" but oral expressions, regional or cultural studies, or, at best, *lingvostranovedenie*, something that Milivojević (1987, 165) describes as the introduction of "non-linguistic aspects relating to the culture of a language" in addition to the teaching of "purely linguistic material".

So, why should sociopragmatics be called sociopragmatics? Or why are other terms misleading? Teaching an FL is about more than mere forms and functions. As the analysis of the coursebook tasks has shown, the main focus lies on (hidden) sociolinguistic appropriateness, propositional precision, coherence and cohesion as well as on the realization of inter- and transactional schemata, but these parts of the construct are never openly disclosed. Flexibility, turn-taking, thematic development and spoken fluency seem to be less important, although they are closely interrelated components. Pragmalinguistics and sociolinguistics need to be interconnected in the FL classroom and need constant attention in the teaching and learning processes. They are essential constituents of the sociopragmatic construct and should not only be taught but also tested and assessed.

6. Sociopragmatic Test Collection Tasks

Having examined how sociopragmatics is dealt with in RFL coursebooks, we will now turn to three test collections offering tasks at A2 and B1 level. We will examine two TORFL (*ТРКИ*) booklets (A2 or "basic level", 2010, and B1 or "first certification" level, 2006), both published by Zlatoust, and a collection of 112 tests in RFL (A1–B1 levels), also published by Zlatoust in 2013.

The TORFL tests normally consist of five parts (reading and writing, listening and speaking, lexis and grammar), offering a narrow selection of test formats, the most common being Multiple Choice (MC).

Sociopragmatic issues are often embedded in writing and speaking tasks, the latter explicitly claim to also test sociolinguistic, discourse, strategic, pragmatic and even *lingvocultural* competences (Balychina 2006, 59-60). The test formats used are O-DCT (mainly for speaking tasks), Writing Notes or Messages, "logical continuation" (for testing reading and discourse or thematic development).

In Example 2 (Tipovye testy [A2] 2010, 25-26), there are two prompts as part of an O-DCT, in which learners are asked either to react to the interlocutor's utterance (a) or to start a conversation (b):

(a) – Вы опоздали на экзамен. Что случилось?
– ...
(b) У вашего друга день рождения. Поздравьте его.
– ...

Example 3 (Tipovye testy [A2] 2010, 63) serves as an illustration of a variety of writing tasks (learners are asked to produce *invitations, excuses, requests,* etc.). Despite the publication year of the test collection (2010), one wonders why the text genre needs to be a hand-written message or note instead of an e-mail or a text message.[5]

Вы пришли к другу, а его нет дома. Напишите ему записку и объясните, зачем вы приходили. В вашей записке должно быть не менее 5 предложений.

Another format that is used in TORFL to test sociopragmatic competence is ST. Example 4 is taken from Balychina (2006, 117):

Расположите реплики в логической последовательности, чтобы восстановить диалог, где один знакомый делится с другим впечатлениями о путешествии из Москвы в Петербург на теплоходе:

1. – Советую тебе поехать отдыхать по тому же маршруту.
2. – Уже вернулся? Понравилась поездка?
3. – Условия на теплоходе что надо: и бытовые, и культурные. Скучно не будет.
4. – Что понравилось больше всего?
5. – Конечно, столько впечатлений!
6. – Я не знаю, куда поехать отдыхать летом.
7. – Я увидел малую Россию, со всеми ее традициями. Углич, Ярославль, Валам, Кижи... [...]
8. – Десять дней плыть по реке. День и ночь... Скучновато...

[5] Cf. the new scale "Correspondence" in the *CEFR Companion Volume* (2018, 94) and the distinction between *notes* from meetings and basic, formal or personal *e-mails*.

Although the test format is mainly used for listening and reading tasks (ibid., 84), this activity not only tests receptive skills, but also the learners' discourse and design competence. The speech acts and interactional schemata implicitly assessed in this task are *ask and react emotionally, talk about one's own impressions, give advice, express skepticism* and *persuade the interlocutor*.[6]

Many TORFL tasks testing speaking, whether at the A2 or B1 level, include sociopragmatic elements. This is revealed by three of the five criteria according to which the performance is assessed (Tipovoj test 2006, 63): understanding the interlocutor's communicative intention; adequate communicative behaviour and achieving the communicative goal in a given situation; following common norms of politeness.

Tsvetova's (2013) test collection provides an ample selection of test formats configured according to the ALTE exam formats (ibid., 6). The skills and competences tested are the same as for TORFL. Even Tsvetova avoids the term "sociopragmatics", but as she asserts in the preface, her test compilation encompasses all the important competences that RFL learners need to acquire–from the linguistic to the communicative to the sociocultural competence. All in all, the 16 tasks are officially assigned to either the writing, listening or speaking skill. The test formats used are O-DCT, MC-DCT, MT, ST and R-WDCT; the test tasks most strongly tapping into the sociopragmatic construct are, alas, disguised as reading tasks.

Example 5 (ibid., 50) is declared as an MC reading task, however, learners need to identify the communicative setting, clarify who the speaking partners are and grasp the communicative intention (e.g. *complaint, encouragement, astonishment*).

Она: – Как не хочется сдавать экономику этой Митрофановой! Говорят, и четвёрку не получить.
Он: – А почему ей? У вас разве не профессор Григорьев читал?
Она: – Да, читал он, а семинары вела Митрофанова …

1. Она, видимо, …
А) студентка
Б) школьница
В) преподаватель

2. Она …
А) боится преподавателя
Б) не понимает преподавателя
В) беспокоится об экзамене

[6] Although the competence level of this task is not given by Balychina, it supposedly tests level B1.

3. Он …
А) удивляется
Б) предлагает сдавать экзамен другому преподавателю
В) думает, что Григорьев – хороший преподаватель
4. Слово «читать» в диалоге значит …
А) «вести лекции»
Б) «вести семинар»
В) «просматривать книги по экономике»

This illustrative overview has shown that sociopragmatic features form an integral part of a number of embedded tasks in RFL test compilations, yet one has to look very carefully, as they are mostly hidden in four-skill and language-in-use tasks. If coursebook authors and test developers were to use the term “sociopragmatic competence” appropriately and consistently, hereby clearly delimiting it from other competences, it would definitely help teachers to see what the purpose of the tasks is–thus becoming more assessment literate.

7. Implications for LAL

The sociopragmatic construct as part of language competence models is never openly disclosed–neither in textbooks, nor in test collections. It often remains superficial (see also Aristizábal 2018) or is marred by other terms such as *lingvostranovedenie* or “intercultural” learning. In textbooks, pragmalinguistic and sociolinguistic competences are often substituted by regional (cultural) topics. In test compilations, they are either concealed in four-skill tasks or embedded in language-in-use tasks. Even if we are inclined to regard such tasks as delivering “integrated skills”, we do not know, if or how the sociopragmatic features figure in the assessment criteria. So, we only partly agree with Krulatz (2014b, 24), who claims that activities focused on (socio)pragmatics “are becoming increasingly common in language classrooms”. This may be true for teaching and learning activities, but it is not true for classroom-based testing. On the contrary: one of the interviewees (K6) said she does not test sociopragmatic characteristics, adding there was no room for this in her lessons.[7] Does this mean her learners’ Russian is sociopragmatically inappropriate and communicatively inadequate? Whether her statement is truly representative of RFL teachers’ opinions in Austria, only further research will show. Yet it suggests that, in her view, a sociopragmatic approach is not vital, nor does she see the interconnectedness between

[7] Interview 6, 31:49–31:58.

linguistic and sociopragmatic competences. As Comer (2012, 152) suggests, "the field may still not be ready to leave mechanical[8] practice behind", but as we know sociopragmatic mistakes are often considered less acceptable by native speakers than grammatical errors (Drackert 2018, 12).

In our opinion, RFL teachers, despite their commitment to a communicative and task-based approach in teaching, still avoid communicative testing and they mostly ignore sociopragmatic issues. We are convinced that teachers need to be made aware of test purposes, of differences between integrated and discrete testing, or of an assessment grid's necessity to reflect the construct in order to evaluate learners' performances "on a basis of valid arguments" (cf. Drackert et al., in this volume).

Furthermore, as long as textbook authors and test developers tuck away sociopragmatic issues in performance tasks and refrain from offering proper assessment criteria, teachers who receive hardly any pre- or in-service training in this field will particularly struggle. As Sickinger and Schneider (2014, 118) also point out, the "sociolinguistic component in the *CEFR* is still critically unspecified" with only one existing scale, "even though it is strongly emphasized in its programmatic view of a communicative linguistic competence" (ibid.). This "un-specification" has also become apparent in our analysis of RFL coursebooks and test collections, although we perfectly understand that as Cohen (2019, 17) puts it "designing items aimed at assessing sociopragmatic awareness effectively is not easy". Nevertheless, we encourage a proper collection of both discrete and integrated activities dedicated to the testing and assessing of sociopragmatic competence which would help to raise RFL teachers' awareness and improve learners' performance. Such a collection of tasks would need to display authentic and interactive materials to be used both in teaching and testing situations.

To also mention a few positive aspects, the variety of test formats encountered in our study is more than satisfactory. Yet a lot is to be done to boost the assessment literacy not only of RFL teachers, but also of textbook authors and test professionals. Presently, Dreher is working on compiling a Russian sociopragmatic test task collection for schools as part of her dissertation project, following Timpe's advice (2012, 116) that an important task to be tackled is the investigation of sociopragmatic comprehension processes before investigating sociopragmatic production.

[8] Mechanical in the sense of "meaning is not essential" and "manipulation of language forms" is more important in RFL exercises, tasks and tests (Comer 2012, 150).

The findings encountered in our analysis of textbooks and test compilations are firm evidence that a sociopragmatic perspective on teaching and testing would accentuate "the importance of the interaction between the individual learner and the social environment" (cf. Leung 2013) in the TL and target culture. On the one hand, it will be the responsibility of textbook authors and test developers to integrate more practical examples of authentic discourse, explicitly reflecting pragmalinguistic and sociolinguistic objectives. Teacher education and subject-specific pedagogy, on the other hand, should promote a higher level of awareness of how language is used sociopragmatically and, as a consequence, how it should be taught, learned and tested in the FL classroom.

Teachers may be well aware of what their students know. But as their teaching programme is often influenced by a (maybe uncritical) use of the textbook and teacher-influenced beliefs of "what is important" in the immediate context, and not so much by the curriculum or a clearly stated construct, teacher-made tests may well tap into an unrepresentative and even invalid construct. And, as teaching input regulates the learning outcome, there might be a mismatch between "local" achievement and "global" proficiency. As teachers need to understand what carelessly developed tasks and non-transparent constructs are and what these may do to the standard of teaching outcomes and test results, it is necessary to increase their expertise by sensitizing them to the fact of "what language is about". The RFL teacher who said that "she should test sociopragmatic competence as that is what is exciting in reality" needs to understand that she would have to introduce sociopragmatics into her teaching in the first place–and this could be exciting, too, and would, above all, result in positive washback.

8. References

Afonin, Sergej. 2013. "Curriculare und methodische Aspekte des Russischunterrichts. Zur Herausbildung der sozio- und pragmatischen Kompetenz im kommunikativ ausgerichteten Russischunterricht." In *Aktuelle Aspekte der Erforschung und Vermittlung der russischen Sprache und Literatur. Materialien der Internationalen wissenschaftlichen Konferenz in Halle (Sachsen-Anhalt, Deutschland). 11.–12. Oktober 2012 (=Das Neue in Erforschung und Vermittlung des Russischen, Bd. 1)*, edited by Sergej Afonin, Swetlana Mengel, and Elena Plaksina, 89-107. Schöneiche bei Berlin: Plaksina Verlag.

Aristizábal, Frank G. 2018. “A Diagnostic Study on Teachers’ Beliefs and Practices in Foreign Language Assessment.” *Medellín, Colombia* 23(1):25-44.

Bachman, Lyle F. 1990. *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, Lyle F., and Adrian S. Palmer. 1996. *Language Testing in Practice: Designing and Developing Language Tests*. Oxford/New York: OUP.

Balychina. 2006 = Балыхина, Татьяна М. 2006. *Основы теории тестов и практика тестирования (в аспекте русского языка как иностранного): Учебное пособие.* Москва: Русский язык. Курсы.

Blum-Kulka, Shoshana. 1982. “Learning to Say What You Mean in a Second Language: A Study of the Speech Act Performance of Hebrew Second Language Learners.” *Applied Linguistics* 3:29-59.

Canale, Michael, and Merrill Swain. 1981. “Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing.” *Applied Linguistics* 1:1-47.

Černaja and Ivanova. 2012 = Черная, Ольга В., and Любовь К. Иванова. “Формирование социопрагматической компетентности студентов-иностранцев в вузе: критерии и показатели эффективности.” *Вестник ТГУ* 1(105):74-78.

Cohen, Andrew D. 2019. “Considerations in Assessing Pragmatic Appropriateness in Spoken Language.” *Language Teaching* 1:1-20. doi:10.1017/S0261444819000156.

Comer, William J. 2012. “Communicative Language Teaching and Russian: The Current State of the Field.” In *Russian Language Studies in North America: New Perspectives from Theoretical and Applied Linguistics*, edited by Veronika Makarova, 133-160. London: Anthem Press.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment (=CEFR)*. Cambridge: Cambridge University Press.

—. 2018. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors.*

Davison, Chris, and Constant Leung. 2009. “Current Issues in English Language Teacher-based Assessment.” *TESOL Quarterly* 43(3):393-415.

Drackert, Anastasia. 2018. “Was würdest du in dieser Situation sagen? Zu Testmöglichkeiten (sozio)pragmatischer Kompetenzen.” *PRAXIS Fremdsprachenunterricht*, *Basisheft* 1:12-14.

Drackert, Anastasia, and Wolfgang Stadler. 2017. “Leistungsbeurteilungskompetenz von Russischlehrkräften in Deutschland, Österreich, der Schweiz und Südtirol (DACHS): Zwischen Status Quo und aktuellen Bedürfnissen.” *Zeitschrift für Fremdsprachenforschung (ZFF)* 28(2): 233-258.

—. 2018. “Assessment Literacy of Russian as a Foreign Language (RFL) Teachers.” In *Current Trends and Future Perspectives in Russian Studies. Актуальные проблемы и перспективы русистики. (=Proceedings of the International Conference on Russian Studies at the University of Barcelona 2018),* edited by Joan Castellví, Andrey Zainouldinov, Ivan Garcia, and Marc Ruiz-Zorrilla, 963-973. Barcelona: Trialba Ediciones.

Drackert, Anastasia, Carmen Konzett-Firth, Wolfgang Stadler, and Judith Visser (in this volume). “An Empirical Study on Romance Language Teachers’ Subjective Theories Regarding Assessment Purposes and Good Tests.” In *Language Assessment Literacy: From Theory to Practice*, edited by Dina Tsagari. Cambridge: Cambridge Scholars Publishing.

Dreher, Anna. in prep. *Testen der soziopragmatischen Komponente kommunikativer Kompetenz im Russischunterricht (Erstellen von Testaufgaben in russischer Sprache auf der Basis einer Analyse ausgewählter Lehrwerke für Russisch als Fremdsprache) (Working Title)*. Dissertation in progress at Innsbruck University.

Erofeeva, Nesterova, and Jurkov. 2017 = Ерофеева, Инна Н., Татьяна Е. Нестерова, and Евгений Е. Юрков. 2017. “Коммуникативная компетенция: структура, содержание, оценка.” In *Актуальные вопросы языкового тестирования,* edited by Ирина Ю. Павловская (выпуск 2), 98-108. СПб: Издательство СПбГУ.

Gesamte Rechtsvorschrift für Lehrpläne–allgemeinbildende höhere Schulen, Fassung vom 01.09.2018. Retrieved from https://tinyurl.com/yb8b3zdy.

Grundy, Peter. 2014. “Review of V. LoCastro’s Pragmatics for Language Educators: A Sociolinguistic Perspective.” *ELT Journal* 68(2):208-211.

Hamann, Carola. 2012. *Otlično A2* (unterstützt von Katharina Berndt, Olga Stelter, Joachim Wiese, Larissa Tielebein, Alla Zobel, Elena Baranova, Andrej Zajarnyj, Oksana Fischer), 1. Auflage. Augsburg: Hueber Verlag.

—. 2018. *Otlično aktuell A1*: *Lehrerhandbuch*. München: Hueber Verlag. Retrieved from https://tinyurl.com/y3a2o78b.

Hamann, Carola, and Irina Augustin. 2014. *Otlično B1*, 1. Auflage. München: Hueber Verlag.

Hymes, Dell H. 1972. "On Communicative Competence." In *Sociolinguistics*, edited by John B. Pride, and Janet Holmes, 269-293. London: Penguin.

Inbar-Lourie, Ofra. 2013. "Guest Editorial to the Special Issue on Language Assessment Literacy." *Language Testing* 30(3):301-307.

Kanik, Mehmet. 2013. "Reverse Discourse Completion Task as an Assessment Test for Intercultural Competence." *Studies in Second Language Learning and Teaching* 3/4: 621-644. Department of English Studies, Faculty of Pedagogy and Fine Arts, Adam Mickiewicz University, Kalisz.

Kapitonova et al. 2006 = Капитонова, Тамара И., Ирина И. Баранова, Елена В. Городецкая, Ольга М. Никитина, Галина А. Плоткина, and Елена В. Жукова. 2006. *Живём и учимся в России*: *Учебное пособие по русскому языку для иностранцев (I уровень)*. Санкт-Петербург: Златоуст.

Kasper, Gabriele, and Kenneth R. Rose, eds. 2001. *Pragmatics in Language Teaching*. Cambridge: Cambridge University Press.

Kolotova and Kofanova. 2012 = Колотова, Наталья И., and Евгения И. Кофанова. 2012. "Формирование прагматической компетенции студентов-филологов в процессе изучения русского языка как иностранного." *Электронный научный журнал Курского государственного университета* 1/21: 378-383.

Kramsch, Claire. 2011. "The Symbolic Dimensions of the Intercultural." *Language Teaching* 44:354-367.

Krulatz, Anna. 2014a. "Electronic Requests in Native and Non-native Russian: Insights into Foreign Language Learners' Sociolinguistic Competence." *Journal of Linguistics and Language Teaching* 5(1):87-97.

—. 2014b. "Integrating Pragmatics Instruction in a Content-based Classroom." *ORTESOL Journal* 31:19-25.

—. 2015. "Judgments of Politeness in Russian: How Non-native Requests Are Perceived by Native Speakers." *Intercultural Communication Studies* XXIV(1): 103-122.

Kusevska, Marija. 2014. "Connecting Development of Pragmatic Competence with the CEFR." *Linguistica* 54(1): 97-112.

Leung, Constant. 2013. "Classroom-based Assessment Issues for Language Teacher Education." In *The Companion to Language Assessment*, edited by Antony J. Kunnan. John Wiley & Sons, Inc. doi:10.1002/9781118411360.wbcla064.

Milivojević, Dragan. 1987. "On the Soviet Interpretation of 'lingvostranovedenie'." *Russian Language Journal Русский язык*, 41 (138/139):165-168.

Ovčinnikova and Kolotova. 2017 = Овчинникова, Майя В., and Наталья И. Колотова. 2017. "Формирование прагматической компетенции иностранных студентов на уроках РКИ как условие успешного коммуникативного взаимодействия." *Ученые записки. Электронный научный журнал Курского государственного университета* 2(42).

Roever, Carsten. 2011. "Testing of Second Language Pragmatics: Past and Future." *Language Testing* 28(4):463-481.

—. 2012. "What Learners Get for Free: Learning of Routine Formulae in ESL and EFL Environments." *ELT Journal* 66(1):10-21.

—. 2014. "Assessing Pragmatics." In *The Companion to Language Assessment*, edited by Antony J. Kunnan, Part 2, 125-139. Chichester: Wiley & Sons.

Scarino, Angela. 2017. "Developing Assessment Literacy of Teachers of Languages: A Conceptual and Interpretive Challenge." *Papers in Language Testing and Assessment* 6(1):18-40.

Sickinger, Pawel, and Klaus Peter Schneider. 2014. "Pragmatic Competence and the CEFR: Pragmatic Profiling as a Link between Theory and Language Use." *Linguistica* 54(1):113-127.

Stadler, Wolfgang. 2015. "Преподавание и тестирование социопрагматической компетенции РКИ." In *Русский язык и литература в пространстве мировой культуры. Материалы XIII Конгресса МАПРЯЛ (Гранада, 13–20 сентября 2015 года),* edited by А. Л. Вербицкая, К. А. Рогова, Т. И. Попова et al., 363-369. Санкт-Петербург: МАПРЯЛ.

—. 2018. "Soziolinguistik und Pragmatik im Bereich „Russisch als Fremdsprache" testen: Die Rolle von russischen MuttersprachlerInnen und HerkunftssprecherInnen bei der Entwicklung von Testitems in russischer Sprache." In *Formen der Mehrsprachigkeit. Sprachen und Varietäten in sekundären und tertiären Bildungskontexten*, edited by Monika Dannerer, and Peter Mauser, 311-332. Tübingen: Stauffenburg.

Sternin. 2013 = Стернин, Иосиф А. 2013. *Анализ коммуникативных ситуаций*. Изд. 2, испр. и доп. Воронеж: Истоки.

Timpe, Veronika. 2012. "Strategic Decoding of Sociopragmatic Assessment Tasks–An Exploratory Think-aloud Validation Study." *Second Language Studies* 30(2) (Spring):109-246.

Timpe Laughlin, Veronika, Jennifer Wein, and Jonathan Schmidgall. 2015. "Defining and Operationalising the Construct of Pragmatic Competence: Review and Recommendations." *Research Report ETS RR–15-06*:1-43. Princeton, NJ: Educational Testing Service. Retrieved from https://tinyurl.com/y5v8o8g5.

Tipovoj test. 2006 = *Типовой тест по русскому языку как иностранному: Первый сертификационный уровень. Второй вариант.* 2006. Москва/Санкт-Петербург: ЦМО МГУ/Златоуст.

Tipovye testy. 2010 = *Типовые тесты по русскому языку как иностранному: Базовый уровень. Варианты.* 2010. Санкт-Петербург: Златоуст.

Tschirner, Erwin, Hermann Funk, and Michael König. 2000. *Schnittstellen. Lehrwerke zwischen alten und neuen Medien*. (=Deutsch als Fremdsprache. Mehrsprachigkeit, Unterricht, Theorie). Berlin: Cornelsen.

Tsvetova, Natalya E. 2013. *112 Tests in Russian as a Foreign Language (Levels A1–B1): A Manual for Foreign Students* (2nd ed.). Sankt Petersburg: Zlatoust.

Widdowson, Henry G. 1983. *Learning Purpose and Language Use.* Oxford: Oxford University Press.

Xu, Yueting, and Galvin T. L Brown. 2016. "Teacher Assessment Literacy in Practice: A Reconceptualization." *Teaching and Teacher Education* 58 (August):149-162.

8. Appendix

Table 1: Sociopragmatic construct according to the *CEFR* (2001, 2018)

Levels	Sociopragmatic Construct								
	sociocultural[9]	sociolinguistic	pragmatic						
			discourse				functional		design
		appropriateness	flexibility	turn-taking	thematic development	coherence and cohesion	spoken fluency	propositional precision	inter- and transactional schemata
A1	everyday living, living conditions, interpersonal relations, values, beliefs and attitudes, body language, social conventions, ritual behaviour	establish basic social contact; *simplest* everyday polite forms; greetings and farewells, introductions, please and thank you, sorry	no descriptor	no descriptor	no descriptor	basic linear connectors	very short, isolated pre-packaged utterances; much pausing	communicate basic information about personal details and needs of a concrete type in a simple way	Interactional: paired sequences of actions: questions/ answers; statement/ agreement, disagreement; request, offer, apology/ acceptance, non-acceptance; greeting, toast/ response Transactional: (goal-oriented) cooperative transactions
A2		handle very short social exchange; everyday polite forms; greeting and address; make and respond to invitations, suggestions and apologies; information exchange, requests, opinions and	expand learned phrases through simple recombinations of their elements; adapt well-rehearsed memorized simple phrases to particular circumstances through limited	ask for attention (use simple techniques to) initiate, maintain and close short and simple face-to-face conversation	give examples using "like" or "for example"; tell a story, describe a simple list of points	simple connectors and most frequently occurring connectors	construct phrases on familiar topics with ease; short exchanges; noticeable hesitation, false starts	communicate what s/he wants to say; simple and direct exchange of limited information on familiar and routine matters; in other situations, message needs	

[9] There are no scales available for sociocultural competence in the *CEFR* (reasons for this are given in Council of Europe 2001, 220), but there is a Section 5.1.1.2 in Chapter 5 ("The user/learner's competences"), which is dedicated to "sociocultural knowledge". In Section 7.2.1, "sociocultural knowledge" is defined as referring to "life in the target community and essential differences between practices, values and beliefs in that community and the learner's own society" (Council of Europe 2001, 158).

		attitudes; socialize; the *simplest* common expressions, following basic routines	lexical substitution					compromising	(e.g. purchase of goods or services)
B1		perform and respond to a wide range of language functions; neutral register; aware of salient politeness conventions and acting appropriately; aware of the differences between customs, usages, attitudes, values and beliefs in the two communities	exploit a wide range of simple language flexibly to express much of what s/he wants; adapt expressions to deal with less routine, even difficult situations	initiate, maintain and close simple face-to-face conversation on familiar topics or topics of personal interest	link elements into a sequence of points	link a series of simple elements into a connected, linear sequence of points	express him-/herself with relative ease; pausing for grammatical and lexical planning and repair is very evident	convey simple, straightforward information of immediate relevance; express the main point; explain the main points in an idea or problem with reasonable precision	

CHAPTER ELEVEN

ASSESSMENT ALIGNED WITH ACTIVE, LIFELONG LEARNING: EVIDENCE FROM CYPRUS

SVIATLANA KARPAVA

Effective educational assessment can improve learning (Shepard 2000; Broadfoot et al. 2001). Assessment should be authentic; for deep, sustainable, complex, and worthwhile learning, it needs to be meaningful, include personal involvement, and relate to real-world scenarios (Torrance 1994; Boud and Hawke 2004; Nicol 2009). Innovative or alternative assessment can eliminate the "backwash effect" of summative assessments or examinations (Watkins et al. 2005). This study investigated assessment and language assessment literacy in secondary- and tertiary-level English language courses in Cyprus. It examined the assessment of learning and assessment for learning and the stages and procedures involved in assessment and learning, i.e. design, preparation, evaluation, feedback, and reaction to feedback.

The participants in this study were 20 university EFL tutors and 64 secondary school English teachers. The study employed a mixed-method approach for data collection and analysis. The teachers were first interviewed and then observed while teaching. They also participated in focus group discussions. The analysis of the data showed that the tutors like the idea of communities of shared practice (Lave and Wenger 1996) and developing into reflective practitioners. There are restrictions in terms of time and the curriculum, but, in general, the tutors are aware of the benefits of formative assessment. They support the idea of active learning that presupposes an active reaction to the feedback. There is a tendency for the tutors, at least at the tertiary level, to be able to self-assess their practice, introduce new ideas into their practice, and become facilitators of the learning process. According to them, it is vital to enhance language assessment literacy and to create an appropriate learning environment that actively involves students in the processes of assessment and peer evaluation, thus promoting the development of critical thinking, problem-solving, and metacognitive and collaborative skills.

1. Introduction

Effective educational assessment helps to improve learning (Broadfoot et al. 2001). Assessment should be student-centred and formative. Tunstall (2000) proposes the idea of communities of shared practice in which tutors, as reflective practitioners, integrate ideas around the theory and practice of assessment (Shepard 2000). As an integral part of teaching, assessment should be transparent, understandable, consistent, and reliable (Radnor and Shaw 1995; Saunders and Davis 1998).

Language assessment training is essential in order to develop language assessment literacy (Taylor 2013). Language assessment literacy (LAL) presupposes the planning, executing, evaluating and reporting of the assessment process and data (Giraldo 2018) as well as test and item analysis (Walters 2010), external test evaluation (Vogt and Tsagari 2014), and the democracy and transparency of assessment practice (Arias et al. 2012), in particular teaching contexts (Scarino 2013). Students need to be aware of assessment methods and criteria and the choice of these methods should be negotiated. Test design, use, interpretation and evaluation are essential for LAL. Teachers should have a critical stance towards language assessment practice, be fair, ethical and aware of the consequences of assessment (Fulcher 2010).

According to Maclellan (2001), there are different forms of alternative assessment and formative assessment. Students need to be involved in assessment, both in terms of goals and criteria, in order to develop critical thinking, problem-solving, and metacognitive and collaborative skills (Wiggins 1989; Linn et al. 1991; Linn and Baker 1996). All information about a student’s achievements should be collected and synthesized. Social contexts, both positive and negative, should be considered. When an assessment is formative, it can have a positive impact on learning. Higher-order skills are measured, increasing the student’s motivation (Wiggins 1989, 1993; Berlak et al. 1992; Garcia and Pearson 1994).

According to Black and Wiliam (1998), formative assessment improves students’ overall knowledge, addresses their needs, and focuses the direction of their studies. The task of an educator is to establish an appropriate learning environment. Wiliam and Thompson (2007) propose five key strategies of formative assessment: the criteria of successful learning should be clear, learning tasks should be effective, learners should be provided with feedback, students should facilitate each other’s learning, and students should be responsible for their own studies. Black et al. (2006) emphasize the importance of a learner’s autonomy.

Boud and Falchikov (2006) suggest that assessment should focus on long-term and future learning. Reflection and self and peer feedback must be incorporated into learning, grading, and assessment (Black and Wiliam 1998). There ought to be a collaborative approach to assessment and feedback between teachers and students; both students and teachers should benefit from the assessment (Hounsell 2003). A student's self-confidence can be fostered; students need to know that they can succeed, and this can be achieved via peer- and self-assessment.

Assessment should be authentic. For deep, sustainable, complex, and worthwhile learning, assessment needs to be meaningful, with personal involvement, and related to the real world (Torrance 1994; Boud and Hawke 2004; Nicol 2009). Educators should consider their students' perceptions when designing assessments, with research and research-like activities being emphasized (Gibbs and Simpson 2004). Innovative or alternative assessment can potentially eliminate the "backwash effect" of summative assessment or examinations (Watkins et al. 2005).

Educators need to challenge their students' preconceptions about assessment and help them transition from surface to deep learning by harnessing their personal interests and concerns and by incorporating purposeful activities that inspire motivation (Trowler and Trowler 2010). For students to become active learners, they need to link learning to practice by exploring, experiencing, and reflecting; education should be based on enquiry and problem-solving with the support and guidance of teachers and tutors (Lombardi 2007; Boud 2009). Assessments need to be discussed; educators should encourage their students to engage in the assessment by evaluating their own learning, self-regulating their progress, and reflecting on their personal development (McDowell 2008).

Ideally, a balance should be struck between summative and formative assessment. While summative assessment is powerful and pervasive, over-focusing on it often leads to detrimental effects on learning. It is common knowledge that students are unwilling to participate in formative activities because they are not usually graded. Rethinking and re-designing summative assessment or frequent summative (feed-forward) assessment can help to change this situation (Rust 2000). It is important to develop the formative potential of summative assessment or learning-oriented assessment with an emphasis on the deep-learning approach (Carless 2007) and a constructive alignment of teaching, learning, and assessment (Biggs and Tang 2007). Summative assessment gives the student only one chance, while formative assessment provides opportunities for the student to improve. Re-balancing summative and formative assessment will create more space for formative assessment and help educators and students to better allocate their time and effort.

Formative feedback is faster and helps students to improve and develop; it is not a product but a process (Hyland 2000; Hounsell 2003; Crook et al. 2006; McDowell et al. 2008; O'Donovan et al. 2008). Any feedback without discussion is inert and retrospective (Nicol 2010). Many students are dissatisfied with feedback, and a significant percentage either do not read it or do not understand it. Feedback thus looks backwards rather than forwards (Chanock 2000; Higgins et al. 2001; Lillis and Turner 2001; Gibbs 2010; Price et al. 2011). If an assessment is seen as a product rather than a process, then the student is a passive dependent (on a teacher) (Black and Wiliam 1998). Thus, it is essential to take a new approach to feedback: an approach that emphasizes the active role of a student (Nicol 2009). A student's reaction to feedback improves and enhances feedback practice policy and the conceptualization of the feedback process (Boud and Falchikov 2007). Feedback can be seen as a cycle of information and reaction (Nicol 2010). High-quality information, as well as the opportunity to act on feedback, should be provided to students; they should follow the process of internalizing, understanding, reflecting, acting, closing gaps, active engagement, and appropriate use of time (Nicol and Macfarlane-Dick 2006; Price et al. 2011). Students need to raise awareness about the purpose of feedback and develop feedback literacy by seeking, recognizing, understanding, and using feedback (Gibbs and Simpson 2004).

Overall, summative assessment can be broken down into smaller, interlinked components. Feedback can be offered in alternative formats (audio-feedback, podcasts or recorded comments, personal communication) and in a group, class, or individualized setting. Feedback should be feed-forward, resource-intensive, less time-intensive, timely and plentiful, staged, and focused on short tasks (Carless 2007; Merry and Orsmond 2008). Students should be given a choice of when to receive feedback: prior to or after the submission. Peers can also be involved in providing quick, useful, and authentic feedback (Nicol and Milligan 2006).

The task of an educator is to encourage students to become self-assessors and capable lifelong learners with the ability to monitor their progress, to self-critique, to independently judge, and to hone reflective and critical habits. Students should become assessors of their own learning (Boud 2000; Carless 2007). They need to be ready for independent work without supervision and guidance, seeking out resources, setting goals, achieving these goals using various strategies, and monitoring the achievement of these goals. Their learning should be planned and organized; they have to be proactive and both generate and process feedback (Boud 1995; Nicol 2009). Focused discussion and self-critique activities help to build assessment communities and develop assessment literacy and culture.

Language assessment literacy is the knowledge that teachers should have about assessment (Berry et al. 2019) relevant to classroom practice (Malone 2011). According to Davies (2008), assessment literacy is composed of skills, knowledge, and principles. Stiggins (1991) and, later, Falsgraf (2005) defined assessment literacy as the ability of a teacher to evaluate students' performance and to improve the teaching/learning process. Pill and Harding (2013), Taylor (2013), and Harding and Kremmel (2016) view assessment literacy as a constellation of skills and competencies with regard to test creation, implementation, and data analysis in an educational setting. Language assessment literacy presupposes critical analysis of the purpose and relevance of the testing tools and the data obtained with their help (Inbar-Lourie 2008). According to Vogt and Tsagari (2014, 377), language assessment literacy is "the ability to design, develop and critically evaluate tests and other assessment procedures, as well as the ability to monitor, evaluate, grade and score assessment on the basis of theoretical knowledge."

Teachers need to promote the importance of self-review and reflection (Haggis 2006) by designing summative assessment to emphasize the role of reflection in learning. Student learning is assessment driven (Brown et al. 1996). Consequently, it is essential that assessment be valid and reliable (Yorke et al. 2000). The goal of education is to develop higher-order skills, critical thinking, and critical reflection skills. Higher-order skills are not always assessed. Students have problems when they are asked to define, contrast, interpret, or critically evaluate. Transmission-based and content-driven teaching methods focus on assessing memorization (Ramsden 1992; Boud 1995). Criterion-referenced assessment and norm-referenced assessment compare students' performances. Students should acquire not only content knowledge but also critical thinking. It is important that the criteria of their assessments be clear and that the teacher's expectations be transparent. Feedback, learning, teaching, and assessment should be aligned (Biggs 1999).

A deep approach to learning presupposes the engagement of students in their assessment; this develops their self-knowledge, metacognition, awareness, and self-regulation (autonomy) and fosters lifelong learning. Students should have a clear understanding of the purpose of the tasks and reward mechanisms (mark or feedback) in order to judge their learning and performance in line with the standards and expectations. The effectiveness of assessment can be monitored with the help of integrative assessment (e-portfolios, blogs, wikis, self-peer-review, self-regulation, and lifelong learning). A shift from a testing culture to an assessment culture (Birenbaum and Docky 1996) presupposes the integration of assessment, teaching, and learning and the active involvement of the learner in his/her assessment (Sambell et al. 1997).

This study investigated the issue of English language assessment and language assessment literacy at the secondary and tertiary levels in Cyprus, including assessment of learning and assessment for learning and the stages and procedures involved in assessment and learning, i.e. design, preparation, evaluation, feedback, and reactions to feedback.

2. Study

2.1 Participants

The participants in the study were 20 university EFL tutors and 64 secondary school English teachers. Their mean age was 27.6 years old, ranging from 21 to 39 years old. All the participants were female. Regarding their level of education, 30 (35.7%) teachers had only a BA in English studies, 34 (40.4%) had earned an MA degree, and 20 (23.9%) had earned a PhD degree. They had between 1 and 10 years of teaching experience (Mean = 4.3 years; SD = 2.65).

2.2 Materials and procedures

A mixed-method approach was implemented for data collection and analysis. Written questionnaires and oral interviews were used for data collection among the EFL teachers. They were interviewed and then observed while teaching. They also participated in focus group discussions. The questionnaires and the interviews had questions on language assessment, formative and summative, assessment of learning and assessment for learning, the stages and procedures involved in assessment and learning, design, preparation, evaluation, feedback and reaction/action on feedback, awareness about innovative and alternative assessment, reflection, teaching practices, the theory and practice of assessment, and peer observation.

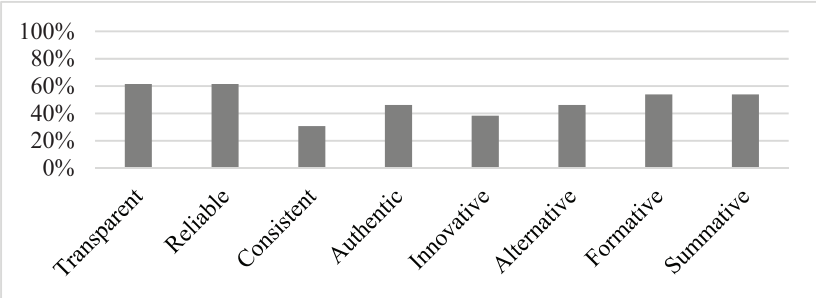
3. Results

The analysis of both the qualitative and quantitative data showed that teachers, both in secondary and tertiary education, have an idea about what constitutes an effective system of assessment. According to them, an effective system of assessment is one that is varied and has multiple and diverse assessment components. It is important for them to be able to gather information on what their students have learned, either through tests and examinations or even via simple oral questions.

Some teachers view assessment as a system that attempts to examine the learner's skills without focusing on only one area, e.g. by testing speaking, listening, reading, and writing skills. Some view assessment as an adjustable system that should, ultimately, meet the needs of the students by including a variety of practices. The system of assessment aims to facilitate a student's improvement far beyond any specific field of study. The participating teachers noted that understanding their students' strengths and weaknesses, helping their students to achieve desirable results, and increasing their students' confidence and motivation are of utmost importance. Assessing a learner means assessing their comprehension and production of written and oral language. In doing so, assessors are required to collect and analyze evidence of improvement in the said skills as well as in their students' abilities to communicate fluently in the L2. When conducting assessments, it is important to emphasize the independent nature of the learners and their ability to make decisions during the acquisition of the L2.

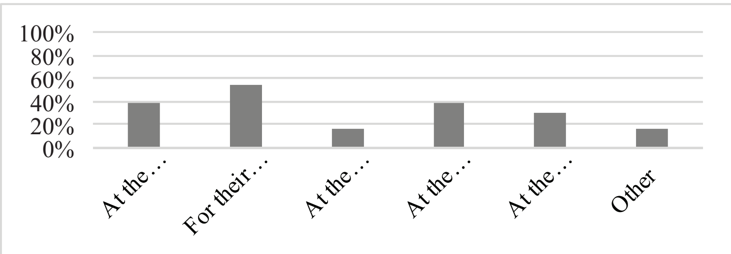
The participants unanimously (100%) stated that assessment could improve learning because, if done correctly, it allows students to expand their knowledge and motivation to learn. In this way, the teacher knows if his/her students have any gaps in knowledge and whether or not they can apply their knowledge (beyond rote memorization). Therefore, the teacher is in a position to know how students use their knowledge and to gauge their readiness and motivation to continue learning. Through assessment, students should be able to show what they understand and to learn from their mistakes. Assessment can facilitate self-realization and improvement provided it is not used as a tool for threats or punishment.

It was revealed that nearly all of the teachers (94%) have time to reflect on their teaching and assessment practice, though there is variation in terms of frequency. They do so throughout the year. Factors such as the curriculum, the number of students, and time pressures negatively affect their ability to reflect, modify assessment components, and provide feedback. Overall, they do give feedback two or three times per term, but ideally, they would like to have time to do it more often. All of the participants, without exception (100%), practise peer observation and peer teaching. This is done roughly one or two times per year. The teachers explain the aims of the studies to their students always (54%), often (38%), or only sometimes (8%). According to the teachers at both levels, the most important characteristics of assessment are transparency and reliability. They implement both formative and summative assessment, as demonstrated in Graph 1.



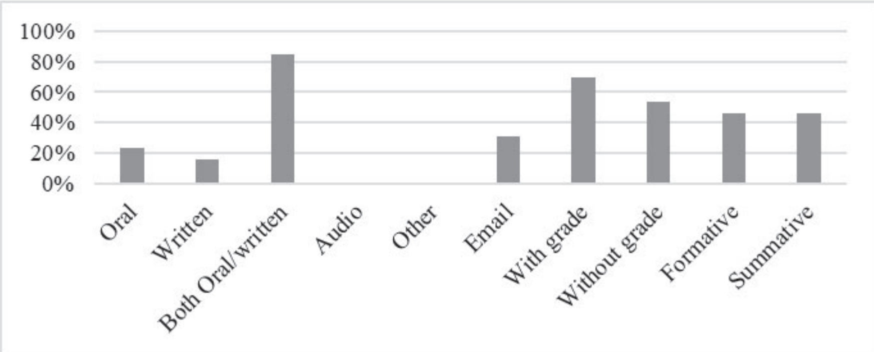
Graph 1: Characteristics of assessment

The teachers always (61%), often (30%), or sometimes (9%) explain the grading criteria to their students. They provide feedback to their students, primarily on their homework, at the end of every session, at the end of the term, or at the end of the year. The details are presented in Graph 2.



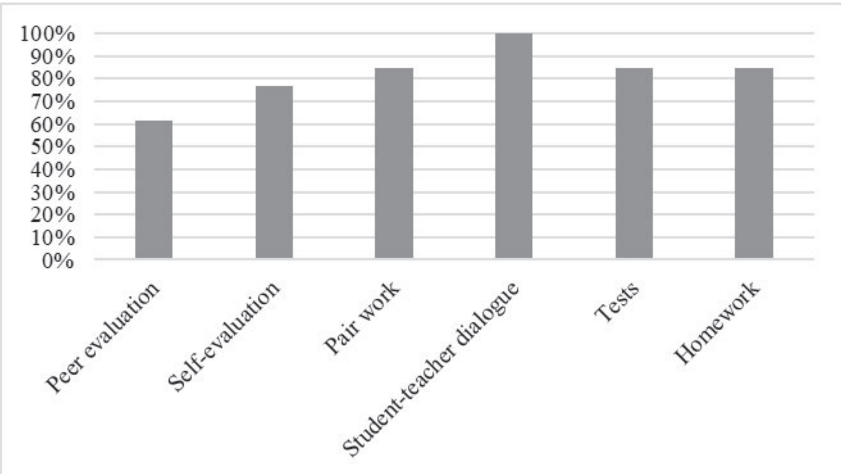
Graph 2: How often do you provide feedback to your students?

It should be noted that the teachers tend to provide both oral and written feedback, normally alongside the marks. They do not provide any audio or online feedback. They implement feedback with both formative and summative assessment, as shown in Graph 3.



Graph 3: What kind of feedback do you provide to your students?

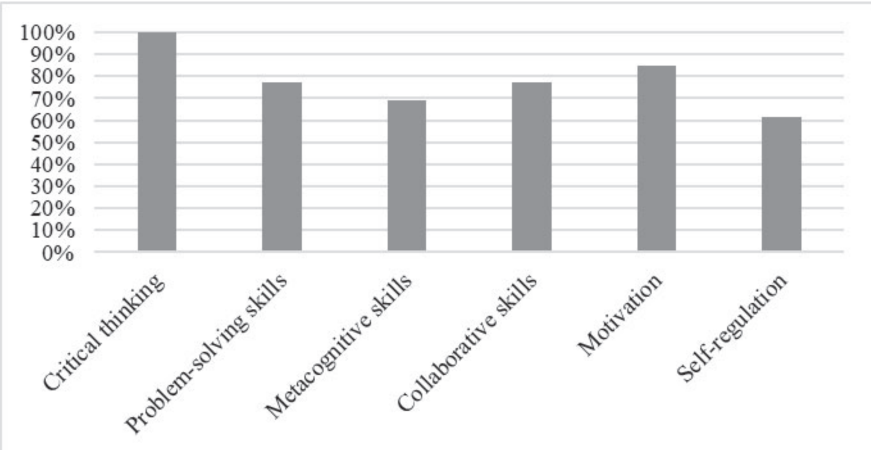
All of the teachers (100%) implement self-assessment, as they think it encourages their students to think critically and gain the independence and motivation needed for learning a language. Self-assessment is implemented when the learners are mature enough to understand their own mistakes. The teachers implement student-teacher dialogues, pair work, self and peer evaluations, tests, and homework, as presented in Graph 4.



Graph 4: Which applies to your classroom teaching and assessment?

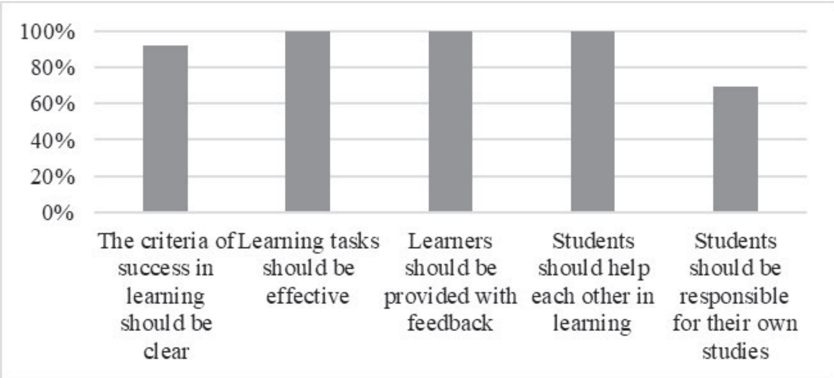
The participants believe that constant development, peer observations, and further training and research can facilitate their professional development and enhance language assessment literacy. Students are able to benefit from the feedback of their peers and identify areas of improvement in their practice. The teachers inform themselves about new methods of education by continuously reading and participating in relevant training. They want to have the freedom to modify their curriculum and assessment components. Constructive discussion with their peers may improve all aspects of their teaching and assessment.

Some of the teachers take notes on what is particularly helpful for their learners, such as an interesting or particularly motivating activity. In addition, they ask their students to share their thoughts on their teaching at least once a month. Many of the teachers attend educational seminars, read recent articles and journals about English language learning and pedagogy, and search recourses online to find new ideas that their learners might like. One important finding was that all of the teachers develop their students' critical skills, as well as their motivation, problem-solving, and collaborative and metacognitive skills, as shown in Graph 5.



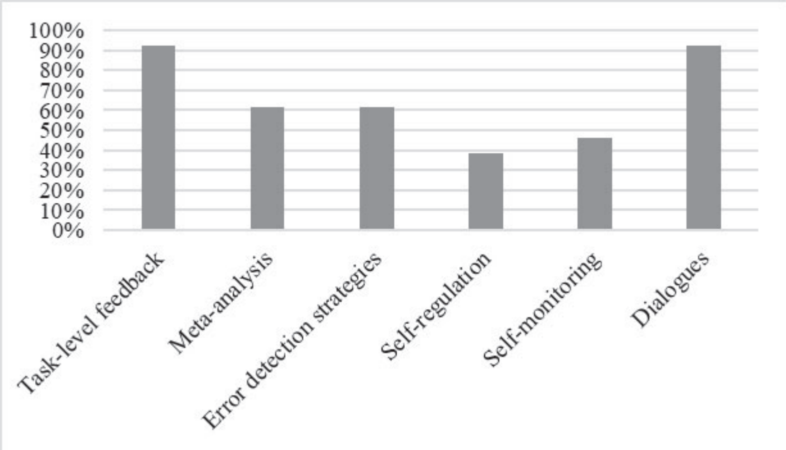
Graph 5: Development of various skills

Not all of the participants were aware of the difference between formative and summative assessment. Those who were aware of the difference provided different definitions of formative assessment, e.g. it takes place throughout the year wherein the students are given feedback but their work is not marked or it helps students to improve their skills in order to do better during summative assessment. The different perspectives given are presented in Graph 6. According to the teachers, it is not an assessment in the traditional sense of having tests at the end of a unit but more like a series of procedures aiming to assess the students' development and identify whether or not changes need to be made in the teaching and learning processes. An alternative assessment is used to provide ongoing feedback to students and monitor their learning. Formative assessment is given throughout the learning period. It is an ongoing process of collecting information in order to assess learners. Regular feedback is not mark driven; rather, it gives an overall picture of a student's performance according to the standard marking criteria of the institution. On the other hand, summative assessment is a formal assessment that usually takes place at the end of the year. This formal assessment evaluates students' progress, using grades, in the areas that have been taught and can take the form of a test, project, or final exam.



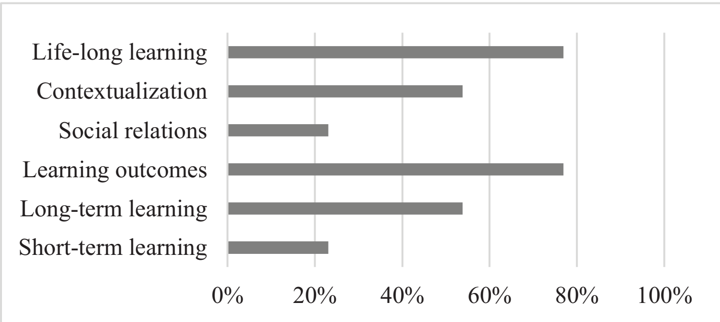
Graph 6: Teachers' views on teaching and learning

The teachers mainly provide task-level feedback and feedback in the format of dialogues, as demonstrated in Graph 7. They stated that they could improve their assessment procedures by taking mixed learning abilities into consideration; in doing so, they would assess each student according to his/her strengths and weaknesses. They believe that with divergence away from summative assessment and incorporation of a more student-focused assessment, learners benefit more, show their true potential, and develop their own learning styles and personalities. Some teachers expressed the view that assessment could be more fun, more straightforward, and less stressful for students if various types of assessment were created, such as oral assessment, written assessment, and group presentations.



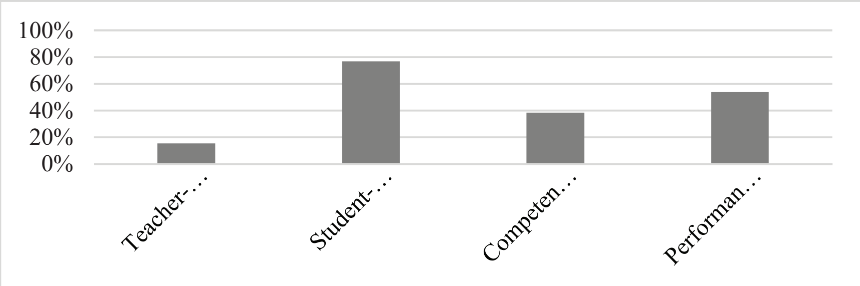
Graph 7: What type of feedback do you provide?

Overall, the participants' assessment focuses on lifelong learning and learning outcomes, as shown in Graph 8. The participants emphasized that there should be a combination of research, practice, and feedback. The teachers should be willing to modify their assessment procedures so that they can provide unbiased and authentic results and so that the students can learn through the assessment process.



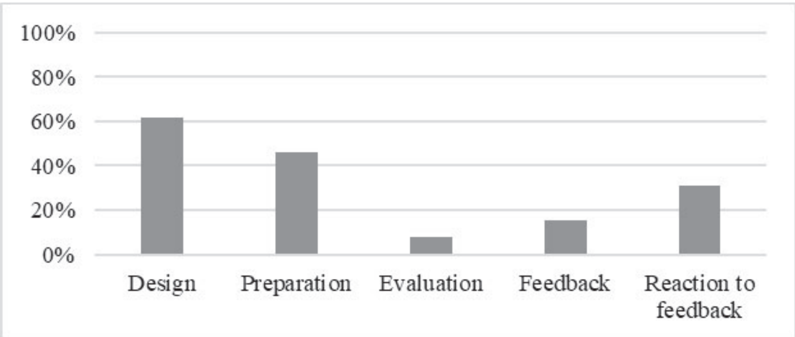
Graph 8: The focus of the assessment

Analysis of the questionnaires, interviews, and focus group discussions revealed that the teachers' assessments are student-focused, as indicated in Graph 9. The teachers in Cyprus hold the view that assessment is practice-oriented and caters to students' needs as well as to their employability. They disagree that assessment at their institutions is authentic, valid, reliable, or capable of reducing anxiety. More than half of the teachers agree that assessment needs to be valid, objective, reliable, transparent, and with explicit criteria. They highlighted the importance of the student's understanding of the assessment criteria in order to improve his/her competencies, self-assessment, self-development, and quality of reflection.



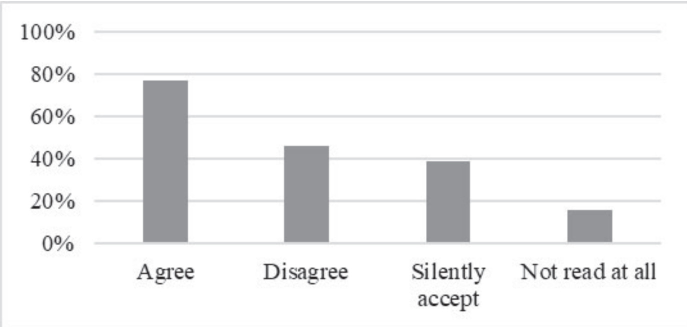
Graph 9: Assessment: characteristics

The teachers stated that more frequent feedback should be provided. Assessment procedures need to be adapted to suit the students' needs. Learners' reflections could be implemented as a criterion. Assessment can be improved by providing feedback, giving extra worksheets and work to do at home for those who are struggling, repeating tasks the learners found difficult, and guiding the students to understand their mistakes via intensive practice in class. According to the participants, the most difficult stages in assessment and learning are design and preparation, as shown in Graph 10.



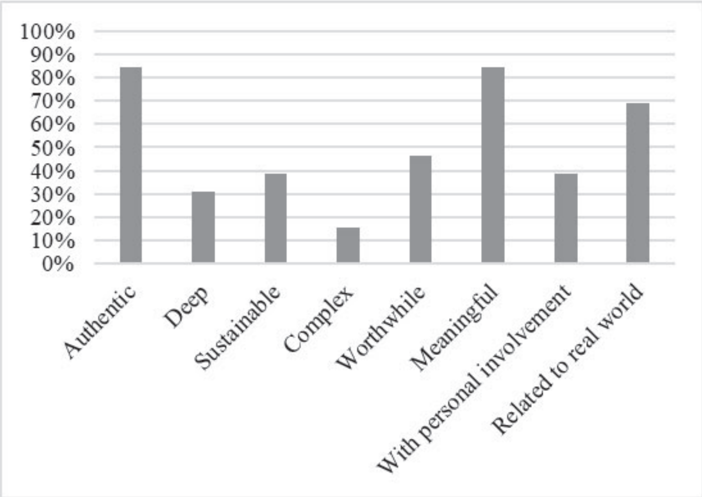
Graph 10: What is the most difficult stage in assessment and learning?

In the majority of cases, students agree with the feedback given, as demonstrated in Graph 11. The teachers consider their students' own reflections on the practice of assessment to be paramount for improvement and for offering the best practice of assessment possible. Through these reflections, the teachers can identify areas that need more attention and tweak the design and style of the assessments based on the students' needs and success. With this student feedback, the teachers understand if they did something wrong and are able to practise more appropriate teaching and assessment methods in the future. Understanding the challenges of being an educator and formulating assessments corresponds to a test of skills that ultimately facilitates the teaching and learning processes.



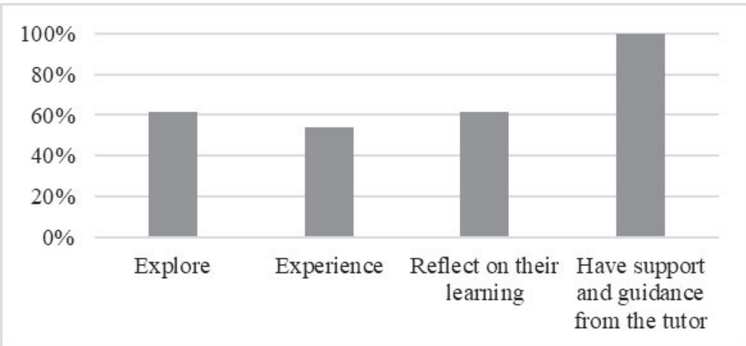
Graph 11: Do your students react to your feedback?

The participants suggested that assessment should be authentic, meaningful, and related to the real world, as presented in Graph 12. It should be noted that both formative and summative assessment are implemented by the teachers, while some of them are unaware of integrative and diagnostic assessment. Around 90% of the respondents claimed that the primary goal of assessment is knowledge construction, and only 40% believe that they need to assess knowledge retention, knowledge transformation, and knowledge control. Memorization of the material is essential for only 10% of the teachers.



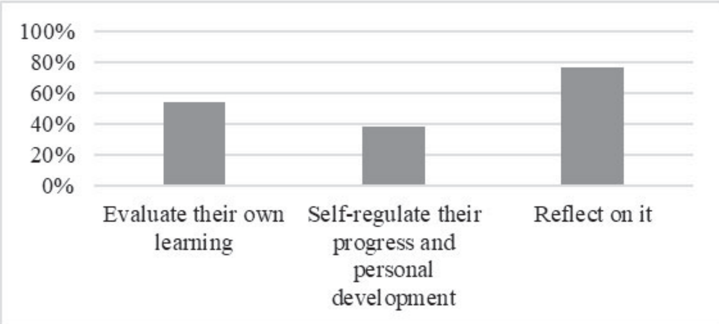
Graph 12: Characteristics of assessment

Most of the learners receive support and guidance from their teachers. Not all of the students are ready to explore, experience, or reflect on their learning, as shown in Graph 13.



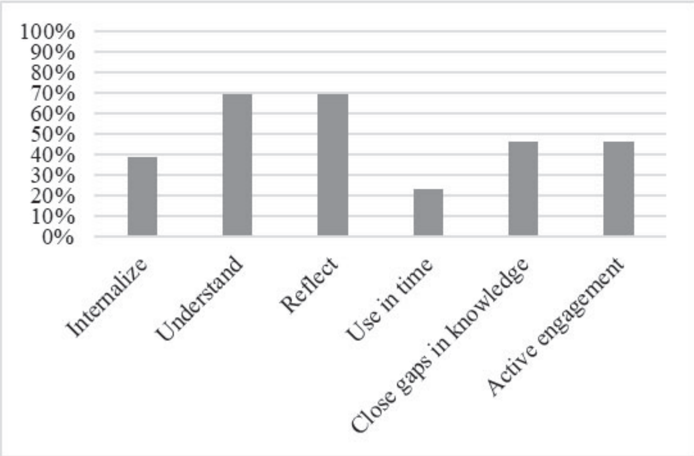
Graph 13: Are your students active learners?

According to the participants, summative assessment can be re-designed, with more tasks and projects being used throughout the year. Concept-checking questions are of great importance. Students are given the opportunity to be active participants in the design of the assessments. Some teachers try to promote a metacognitive approach in order to help their students learn by themselves. They agree that more realistic grading criteria should be implemented. The teachers have considered incorporating innovative elements that are learner-focused in that the students are given the opportunity to show how they apply their knowledge and demonstrate any broader sets of skills/learning styles. Therefore, the focus would shift towards evaluating the students' learning with consideration of their individual strengths. The analysis of the data shows that the teachers encourage personal engagement with students, as demonstrated in Graph 14.



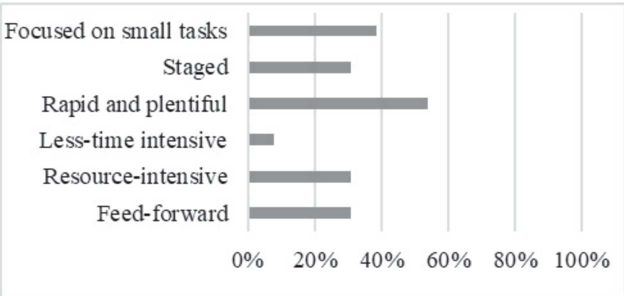
Graph 14: Do you encourage the personal engagement of students?

It was found that the majority of the students study primarily for exams (86%). Most of the teachers (92%) see feedback as a process rather than a product (8%). Feedback is valuable, as are responses to the feedback. The teachers expect their students to understand the assessment and to reflect on it, as presented in Graph 15. Overall, the participants agree that they need to minimize testing and pay attention to the individual needs of each student in order to eliminate a washback effect. In order to do this, the teachers suggested having either more time or less to cover on the curriculum so as to encourage innovative and motivational methods. With more time or less to cover on the curriculum, they would not have to rush and could dedicate time to activities that are more likely to help students develop useful skills and cater to different learning styles. Some participants strongly believe that they should be teaching for the sake of learning rather than teaching solely for the sake of exam pass rates. They highlighted the importance of the learning process rather than focusing on results. Some of the participants admitted that washback is both an inevitable and currently unsolved problem, although they have tried not to let it influence their teaching.



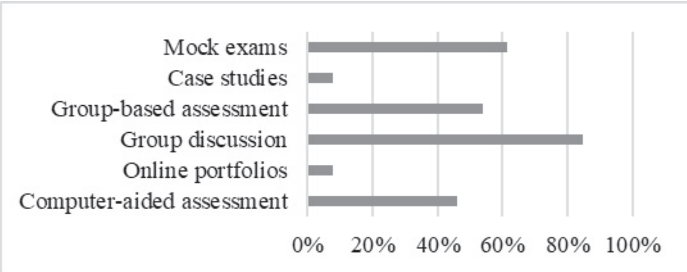
Graph 15: What does action on feedback mean?

The teachers mainly use individualized feedback, group/class feedback, or personal communication over alternative types of feedback, such as recorded comments, podcasts, and audio feedback. The participants provide quick and plentiful feedback, and they employ small tasks, staged, feed-forward, and resource-intensive feedback, as shown in Graph 16.



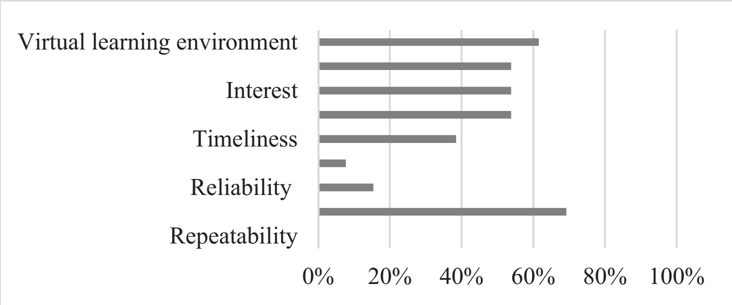
Graph 16: What kind of feedback do you provide?

The type of feedback consists of explanations, motivation and developmental guidance, corrections, and self-regulation. The teachers consider their assessments to be valid and reliable. In their assessments, they mostly take learning outcomes (91%) into consideration rather than course design. Group discussion, mock exams, group-based assessments, and computer-aided assessments are used more frequently than case studies and online portfolios, as demonstrated in Graph 17. The students generally have different reactions to negative feedback. Most of them realize what went wrong, but some do not react positively. Sometimes reactions are initially negative but change once the students understand that the feedback is for their own good. As stated by the teachers, the younger students sometimes feel sad, whereas the older students might become angry or think that the feedback is unfair. They may ask for clarification or indicate that they somehow knew the area that needs improvement, but it depends on the student. Some students do not want to hear negative feedback, while others feel disappointed; some of them react negatively, while others try to improve. Some of the teachers try to be thorough in their written and oral feedback so that the learners understand the reasoning behind any negative feedback, and this usually elicits a positive reaction from the student.



Graph 17: Types of assessment

The participants view computer-aided assessment as an important form of assessment because a virtual learning environment allows for obtaining immediate responses and marks, which is flexible and increases their students' motivation and interest in their studies. This can be seen in Graph 18. In this environment, the students tend to have varied reactions to positive feedback. Overall, they appreciate it. The younger students become excited and sometimes neglect to read all of the details of the feedback. The older students also become excited, but some of them read the feedback in full to check what they did well, which in turn increases their confidence and helps them to understand what is expected of them. According to the participants, positive feedback is always a good asset for students because it increases their confidence. They feel proud of their achievements, smile, and might do their best in subsequent tasks. The teachers noted that students often fail to respond to positive comments, speculating that the students might be more indifferent to positive than to negative feedback. One of the teachers admitted that her students do not usually ask her to provide more explanation, although she still does for the sake of their development.



Graph 18: Why is computer-aided assessment important?

Nearly all of the teachers stated that continuous professional development (CPD) is important. CPD includes the following: the lifelong development of language learning and pedagogy; reflecting on teaching practices and updating resources; the ability to mature through independence in relation to how they deliver modules and assess students; learning from their own mistakes, constantly seeking self-improvement, and accepting criticism as well as praise. The participants believe that they need to learn from their environment by keeping up to date with new methods and attending relevant conferences and exhibitions in order to enhance their language assessment literacy. Continuous learning, broadening one's horizons, and adopting effective practices also contribute to CPD. More than 80% of the respondents claimed that their students care more about their grade than learning (39%), professional development (40%), knowledge (18%), or the curriculum (10%). Depending on the abilities of the students and the type of work, the teachers spend varying amounts of time on marking and providing feedback per student. It ranges from 5 to 45 minutes.

There was no unanimous opinion on whether the teachers' students are active learners (60%) or passive recipients of knowledge (40%). The teachers think that assessment is for learning and of learning (78%). The participants noted the following challenges: time pressures (82%), number of students (53%), heavy workload (51%), students' motivation (48%), and resource constraints (30%). The teachers can enhance CPD via the active involvement of their students in building knowledge (70%), being reflective practitioners (78%), fruitful teacher-student relationships (76%), promoting deep knowledge (62%), harbouring a responsibility for learning and teaching (46%), professionalism (46%), encouraging employability skills (37%), content/knowledge transfer (35%), generic skills (31%), and higher-order skills (24%).

4. Discussion and Conclusion

Analysis of the data indicated that the teachers like the idea of communities of shared practice (Lave and Wenger 1996) and being reflective practitioners. There are restrictions in terms of time and the curriculum, but, in general, they are aware of the benefits of formative assessment. They support the idea of active learning that presupposes an active reaction to feedback. There is an opportunity for the teachers, at both the secondary and tertiary levels, to self-assess their practice, introduce new ideas into their practice, and become facilitators of the learning process. According to the teachers, it is crucial to create an appropriate learning environment that actively involves students in the processes of assessment and peer evaluation, which promote critical thinking, problem-solving, and metacognitive and collaborative skills.

Overall, the teachers aim to broaden their knowledge, become better professionals, and improve their teaching and assessment methods. More formal training on assessment literacy is needed so that the teachers' assessments are not based solely upon their past personal school experience and in-service teaching experience (Smith et al. 2014; Vogt and Tsagari 2014). There should be a close link between assessment and classroom practice (McKay 2006). Both practice- and theory-oriented training can help to boost teachers' confidence with respect to assessment.

Some of the teachers admitted that, with regard to assessment, they focus primarily on summative assessment or exams and their teaching is overly teacher-centred. They are trying to pay more attention to the teacher-student relationship and formative assessment. Research-oriented teaching approaches are implemented in order to develop students' knowledge, including both generic and higher-order skills. Being a teacher is not only a matter of transferring content to the students but of helping them to become active learners and to build knowledge independently. It is important to make the criteria of feedback explicit to the students and to align teaching, learning, and assessment.

Teachers are responsible for their students, but they also need to be responsible for their own learning. Peer observations, communication with colleagues, special seminars, and conferences are fruitful because educators can share problematic issues, reflect, and express their feelings and attitudes with respect to their teaching practices.

At the beginning of their teaching careers, the participants focused on testing their students' passive reproduction of information (a surface approach to learning) rather than understanding, critical thinking, and integration of information (a deep approach to learning). With more years of teaching experience, they are becoming reflective practitioners and are gradually beginning to teach their students accordingly.

This chapter investigated the role of assessment and feedback in the educational process and the ways in which students learn. It emphasized the role of formative assessment and the development of the students' higher-order skills, such as critical thinking, self-assessment, reflection, metacognition, and collaboration. The criteria of assessments should be explicit for students. Students should be actively involved in building their knowledge, with help, guidance, and support from their teachers. Educators should be open to innovative and alternative forms of assessment that foster lifelong learning and are closely related to the students' reflection and peer responses. Assessment is a mutually beneficial process between a teacher and a student. Teachers provide students with feedback, which not only helps them to learn but allows for the evaluation of the teachers' methods. Assessment of learning for learning and feedback is part of the educational cycle. Formative assessment is more efficient for students than summative assessment, as it helps students to improve and progress throughout the year rather than only at the end of the year.

The participants would like to eliminate the backwash effect of the summative assessment and any hidden curriculum strategy. The teachers admitted that the majority of their students study only for tests. Most of the teachers want to change this mentality and re-balance summative and formative assessment, creating more space for formative assessment than before. They are sure that this will stimulate their students' development, stimulate a growth in knowledge, motivation, self-regulation, and reflection, and promote an in-depth approach to learning. It is important to create learning-oriented assessment accompanied by timely, qualitative, and staged feedback. Although it is time-consuming, the teachers will ultimately be rewarded by their students' success and progress. The participants need to consider how to engage their students with feed-forward, individualized feedback, creating a formative learning environment for their students. They should decrease their content-driven teaching and assign progressively challenging tasks, promote active participation, and minimize the hierarchical relationship between students and teachers.

Most of the teachers who took part in this research, at both the secondary and tertiary levels of education, try to encourage their students to engage with their feedback, to discuss it, to react to it, to conceptualize it, and to process and internalize it. They also need to work on other, alternative formats of feedback provision to increase feedback efficiency (e.g. audio files, exemplars, dialogues, discussions), taking into consideration students' perceptions and expectations. Students themselves should be able to monitor, evaluate, and reflect on their studies and progress. They should grow into autonomous and independent learners. Educators need to pay attention to the validity and reliability of their assessment methods. They have to be more careful with group work and joint projects due to the problem of "free-riding." Thus, educators need to rely not only on their judgement but also on peer assessment. Students' evaluation of their teaching practice is of great importance as the teachers try to implement their suggestions and to improve.

Overall, it is of great importance to improve teaching practices and assessment strategies by providing flexible assessment formats in line with students' needs. Awareness about language assessment literacy and teacher training in assessment are crucial (Fulcher 2012; Crusan et al. 2016). A holistic view of assessment is needed rather than the prevalent focus on assessment techniques, ready-made assessment materials, compensation strategies, and past learning/teaching experience. The knowledge obtained through this study, which investigated teachers' practices, beliefs, and attitudes towards assessment, their relatively low level of language assessment literacy, their lack of confidence about engaging in assessment and testing activities (Clapham 2000), and the impact of assessment on learning outcomes, can help to develop comprehensible and practical training material on language assessment literacy that meets actual teachers' needs (Berry et al. 2019), especially in the setting of Cyprus. Among the teachers' goals are the following: to align teaching with students' learning and assessment, to promote the employability of their students, and to ensure that the students are given explicit and transparent assessment criteria and formative feedback. To enhance language assessment literacy, further research is needed on how to increase the efficiency of teaching and learning, taking into consideration dynamic contextual assessment requirements and personal professional development needs (Tsagari and Vogt 2017). Action research can be undertaken in order to improve teaching and assessment methods at schools and universities.

5. References

Arias, Clara Inés, Maturana, Liliana and María Isabel Restrepo. 2012. “Evaluación de los aprendizajes en lenguas extranjeras: hacia practices justas y democráticas” [Evaluation in foreign language learning: Towards fair and democratic practices]. *Lenguaje* 40(1):99-126.

Berlak, Harold, Fred Newmann, Elizabeth Adams, Doug Archbald, Tyrell Burgess, John Raven, and Thomas Romberg. 1992. *Towards a New Science of Educational Testing and Assessment*. New York: University of New York Press.

Berry, Vivien, Susan Sheehan, and Sonia Munro. 2019. “What Does Language Assessment Literacy Mean to Teachers?” *ELT Journal* 73(2):113-123.

Biggs, John, and Catherine Tang. 2007. *Teaching for Quality Learning at University*. Maidenhead: Open University Press.

Biggs, John. 1999. *Teaching for Quality Learning at University. Buckingham: Society for Research into Higher Education*. Buckingham: Open University Press.

Birenbaum, Menucha, and Filip Dochy. 1996. *Alternatives in Assessment of Achievements, Learning Processes and Prior Knowledge*. Dordrecht: Kluwer.

Black, Paul, and Dylan Wiliam. 1998. “Assessment and Classroom Learning.” *Assessment in Education* 5(1):7-74.

Black, Paul, Robert McCormick, Mary James, and David Pedder. 2006. “Learning how to Learn and Assessment for Learning: A Theoretical Inquiry.” *Research Papers in Education* 21(2):119-132.

Boud, David, and Geof Hawke. 2004. *OVAL. Research Working Paper* 03-17, Sydney. The Australian Centre for Organisational, Vocational and Adults Learning (OVAL), University of Technology, Sydney.

Boud, David, and Nancy Falchikov. 2006. “Aligning Assessment with Long-term Learning.” *Assessment and Evaluation in Higher Education* 31(4):399-413.

—. 2007. *Rethinking Assessment in Higher Education: Learning for the Longer Term*. London: Routledge.

Boud, David. 1995. *Enhancing Learning through Self-assessment*. London: Kogan Page.

—. 2000. “Sustainable Assessment: Rethinking Assessment for the Learning Society.” *Studies in Continuing Education* 22(2):151-167.

—. 2009. “How can Practice Reshape Assessment?” In *Assessment, Learning and Judgment in Higher Education*, edited by Gordon Joughin, 29-43. Dordrecht: Springer.

Broadfoot, Patricia, Marilyn Osborn, Keith Sharpe, and Claire Planel. 2001. "Pupil Assessment and Classroom Culture: A Comparative Study of the Language of Assessment in England and France." In *Curriculum and Assessment*, edited by David Scott, 41-62. Westport, CT: Ablex Publishing.

Brown, Sally, Phil Race, and Brenda Smith. 1996. *500 Tips on Assessment*. London: Kogan Page.

Carless, David. 2007. "Conceptualizing Pre-emptive Formative Assessment." *Assessment in Education: Principles, Policy and Practice* 14(2):171-184.

Chanock, Kate. 2000. "Comments on Essays: Do Students Understand What Tutors Write?" *Teaching in Higher Education* 5(1), 95-105.

Clapham, Caroline. 2000. "Assessment and Testing." *Annual Review of Applied Linguistics* 20:147-61.

Crook, Charles, Harriet Gross, and Roy Dymott. 2006. "Assessment Relationship in Higher Education. The Tension of Process and Practice." *British Educational Research Journal* 32(1):95-144.

Crusan, Deborah, Lia Plakans, and Atta Gebril. 2016. "Writing Assessment Literacy: Surveying Second Language Teachers' Knowledge, Beliefs, and Practices." *Assessing Writing* 28:43-56.

Davies, Alan. 2008. "Textbook Trends in Teaching Language Testing." *Language Testing* 25(3):327-47.

Falsgraf, Carl. 2005. "Why a National Assessment Summit? New Visions in Action. National Assessment Summit." Meeting Conducted in Alexandria, Va. Retrieved April 11, 2017. Retrieved from: http://www. nflrc.iastate.edu /nva/word documents/ assessment_2005/ pdf /nsap_introduction.pdf.

Fulcher, Glenn. 2010. *Practical Language Testing*. London, UK: Hodder Education.

—. 2012. "Assessment Literacy for the Language Classroom." *Language Assessment Quarterly* 9(2):113-32.

Garcia, Georgia, and P. David Pearson. 1994. "Assessment and Diversity." *Review of Research in Education* 20(1):337-391.

Gibbs, Graham, and Claire Simpson. 2004. "Conditions under which Assessment Supports Students' Learning." *Learning and Teaching in Higher Education* 1(1):3-31.

Gibbs, Graham. 2010. *Using Assessment to Support Students' Learning*. Leeds: Leeds Metropolitan University.

Giraldo, Frank. 2018. "Language Assessment Literacy: Implications for Language Teachers." *Profile: Issues in Teachers' Professional Development* 20(1):179-195.

Haggis, Tamsin. 2006. “Pedagogies for Diversity: Retaining Critical Challenge amidst Fears of ‘Dumbing down’.” *Studies in Higher Education* 31(5):521-535.

Harding, Luke, and Benjamin Kremmel. 2016. “Teacher Assessment Literacy and Professional Development.” In *Handbook of Second Language Assessment. Handbooks of Applied Linguistics*, vol. 12, edited by Dina Tsagari, and Jayanti Banrjee, 413-428. Berlin: Mouton De Gruyter.

Higgins, Richard, Peter Hartley, and Alan Skelton. 2001. “Getting the Message Across: The Problem of Communicating Assessment Feedback.” *Teaching in Higher Education* 6(2):269-74.

Hounsell, Dai. 2003. “Student Feedback, Learning and Development.” In *Higher Education and the Lifecourse*, edited by Maria Slowey, and David Watson, 67-78. Buckingham: SRHE/Open University Press.

Hyland, Paul. 2000. “Learning from Feedback on Assessment.” In *The Practice of University History Teaching*, edited by Paul Hyland, and Alan Booth, 233-247. Manchester: Manchester University.

Inbar-Lourie, Ofra. 2008. “Constructing a Language Assessment Knowledge Base: A Focus on Language Assessment Course.” *Language Testing* 25(3):385-402.

Lave, Jean, and Etienne Wenger. 1996. “Legitimate Peripheral Participation in Communities of Practice.” In *Learning and Knowledge*, edited by Robert McCormick, and Carrie Paechter, 21-35. London: Paul Chapman.

Lillis, Theresa, and Joan Turner. 2001. “Student Writing in Higher Education: Contemporary Confusion, Traditional Concerns.” *Teaching in Higher Education* 6(1):57-68.

Linn, Robert, and Eva Baker. 1996. “Can Performance-based Student Assessments be Psychometrically Sound?” In *Performance-based Student Assessment: Challenges and Possibilities,* Ninety-Fifth Yearbook, edited by Joan Baron, and Dennie Wolf, 84-103. Chicago, IL: National Society for the Study of Education.

Linn, Robert, Eva Baker, and Stephen Dunbar. 1991. “Complex Performance-based Assessment: Expectations and Validation Criteria.” *Educational Researcher* 20(8):15-21.

Lombardi, Marilyn. 2007. *Authentic Learning for the 21st Century: An Overview*, ELI Paper I. Educause. Online Available: http://www.educause.edu.

Maclellan, Effie. 2001. “Assessment for Learning: the Different Perceptions of Tutors and Students.” *Assessment and Evaluation in Higher Education* 26(4):307-318.

Malone, Margaret. 2011. "Assessment Literacy for Language Educators." *CAL Digest October 2011*. Available at www.cal.org.

McDowell, Liz. 2008. "Negotiating Assignment Pathways: Students and Academic Assignments." *Teaching in Higher Education* 13(4):423-435.

McKay, Penny. 2006. *Assessing Young Language Learners*. Cambridge: Cambridge University Press.

Merry, Stephen, and Paul Orsmond. 2008. "Students' Attitudes to a Usage of Academic Feedback Provided via Audio Files." *Bioscience Education* 11(1):1-11.

Nicol, David, and Collin Milligan. 2006. "Rethinking Technology-supported Assessment in Terms of the Seven Principles of Good Feedback Practice." In *Innovative Assessments in Higher Education,* edited by Karen Clegg, and Cordelia Bryan, 64-78. London: Routledge.

Nicol, David, and Debra Macfarlane-Dick. 2006. "Formative Assessment and Self-regulated Learning: A Model and Seven Principles of Good Feedback Practice." *Studies in Higher Education* 31(2):199-218.

Nicol, David. 2009. "Assessment for Learner Self-regulation: Enhancing Achievement in the First Year Using Learning Technologies." *Assessment and Evaluation in Higher Education* 34(3):335-52.

—. 2010. "From Monologue to Dialogue: Improving Written Feedback Processes in Mass Higher Education." *Assessment and Evaluation in Higher Education* 35(5):501-517.

O'Donovan, Berry, Margaret Price, and Chris Rust. 2008. "Developing Student Understanding of Assessment Standards: A Nested Hierarchy of Approaches." *Teaching in Higher Education* 13(2):205-217.

Pill, John, and Harding, Luke 2013. "Defining the Language Assessment Literacy Gap: Evidence from a Parliamentary Enquiry." *Language Testing* 30(3):381-402.

Price, Margaret, Karen Handley, and Jill Miller. 2011. "Feedback: Focusing Attention on Engagement." *Studies in Higher Education* 36(8):879-896.

Radnor, Hilary, and Ken Shaw. 1995. "Developing a Collaborative Approach to Moderation." In *Evaluating Authentic Assessment. Buckingham*, edited by Harry Torrance, 124-143. Open University Press.

Ramsden, Paul. 1992. *Learning to Teach in Higher Education*. London: Routledge.

Rust, Chris. 2000. "A Possible Student-centered Assessment Solution to Some of the Current Problems of Modular Degree Programmes (opinion piece)." *Active Learning in Higher Education* 1(2):126-131.

Sambell, Kay, Liz MacDowell, and Sally Brown. 1997. "But is it Fair?: An Exploratory Study of Student Perceptions of the Consequential Validity of Assessment." *Studies in Educational Evaluation* 23(4): 349-71.

Saunders, Mark, and Susan Davis. 1998. "The Use of Assessment Criteria to Ensure Consistency of Marking: Some Implications for Good Practice." *Quality Assurance in Higher Education* 6(3): 162-171.

Scarino, Angela. 2013. "Language Assessment Literacy as Self-awareness: Understanding the Role of Interpretation in Assessment and in Teacher Learning." *Language Testing* 30(3):309-327.

Shepard, Lorrie. 2000. "The Role of Assessment in a Learning Culture." *Educational Researcher* 29(7):1-14.

Smith, Lisa, Mary Hill, Bronwen Cowie, and Alison Gilmore. 2014. "Preparing Teachers to Use the Enabling Power of Assessment." In *Designing Assessment for Quality Learning*, edited by Claire Wyatt-Smith, Valentina Klenowski, and Peta Colbert, 303-23. Dordrecht: Springer.

Stiggins, Richard. 1991. "Assessment Literacy." *Phi Delta Kappan* 72(7):534-539.

Taylor, Lynda. 2013. "Communicating the Theory, Practice and Principles of Language Testing to Test Stakeholders: Some Reflections." *Language Testing* 30(3): 403-412.

Torrance, Harry. 1994. *Evaluating Authentic Assessment: Problems and Possibilities in New Approaches to Assessing*. Buckingham: Open University Press.

—. 2007. "Assessment as Learning? How the Use of Explicit Learning Objectives, Assessment Criteria and Feedback in Post-secondary Education and Training can Come to Dominate Learning." *Assessment in Education Principles, Policy and Practice* 14(3):281-294.

Trowler, Vicki, and Paul Trowler. 2010. *Student Engagement Evidence Summary.* York Higher Education Academy. Online. Available <http//www.heacademy.ac.uk>.

Tsagari, Dina, and Karin Vogt. 2017. "Assessment Literacy of Foreign Language Teachers around Europe: Research, Challenges and Future Prospects." *Papers in Language Testing and Assessment* 6(1):41-63.

Vogt, Karin, and Dina Tsagari. 2014. "Assessment Literacy of Foreign Language Teachers: Findings of a European Study." *Language Assessment Quarterly* 11(4):374-402.

Walters, F. Scott. 2010. "Cultivating assessment literacy: Standards evaluation through language-test specification reverse engineering." *Language Assessment Quarterly* 7(4):317-342.

Watkins, David, Bo Dahlin, and Mats Edcholm. 2005. "Awareness of the Backwash Effect of Assessment: A Phenomenographic Study of the Views of Hong Kong and Swedish Lecturers." *Instructional Sciences* 33(4):283-309.

Wiggins, Grant. 1989. "Teaching to the (Authentic) Test." *Educational Leadership* 46(7):41-47.

—. 1993. *Assessing Student Performance*. San Francisco, CA: Jossey-Bass.

Wiliam, Dylan, and Marnie Thompson. 2007. "Integrating Assessment with Instruction: What will it Take to Make it Work?" In *The Future of Assessment: Shaping Teaching and Learning,* edited by Carol Dwyer, 53-82. Mahwah, NJ: Erlbaum.

Yorke, Mantz, Paul Bridges, and Harvey Woolf. 2000. "Mark Distributions and Marking Practices in UK Higher Education." *Active Learning in Higher Education* 1(1):7-27.

CHAPTER TWELVE

DESIGNING AN ASSESSMENT ENVIRONMENT FOR GIFTED STUDENTS: FOCUS ON TEACHER EDUCATION

MARINA PEREVERTKINA

The expectations of teachers and their professional activities are changing, and new professional standards are being implemented. An analysis of these standards among various countries reveals wide variations in their functions and application, though all of them define teacher competences in terms that are both country-specific and universal. One of these universal metrics of professional competence is language assessment literacy defined as the know-how to obtain the information required to make pedagogical decisions about students, to give feedback, to judge instructional effectiveness and curricular adequacy, and to inform policy. Given the gap between the current methodologies and the standard requirements, the development of competence in this area has become a challenge in teacher training.

Another increasing concern is the education, guidance, and assessment of gifted students who possess a greater knowledge base in foreign languages and are capable of understanding and retrieving information faster, responding to tasks more efficiently, and selecting better problem-solving strategies. The paucity of curricula and courses aimed at developing an educator's ability to assess gifted students has motivated this research. The chapter is intended to serve two main purposes, namely to clarify the environment required for accurately assessing gifted students and to outline the strategies to be used in choosing valid techniques of assessment, developing the assessment methods appropriate for gifted students, and scoring and interpreting the results of those assessments such that they may be efficiently used by teachers of English when planning their curricula.

The chapter will also review the content of a course that was designed at Herzen State Pedagogical University of Russia specifically to develop the language assessment literacy of future teachers of English and train them to design an environment of assessment for gifted students.

The results of this review indicate that the quality of teacher training will be enhanced by the methods proposed here, which will enable future teachers of

English to meet today's new standards, while gifted students in the gifted cluster classes will subsequently experience greater academic growth.

1. Introduction

Teaching is a dynamic profession and with the implementation of new professional standards the expectations of teachers and their professional activities are changing. An analysis of related documents from the European Commission (i.e., "Supporting the Teaching Professions" and "Supporting Teacher Competence Development"); of UK Teachers' Standards; of the National Standards for Foreign Language Education (USA); and of the Russian Teachers' Professional Standard, while revealing wide variations in their functions and application, define teachers' competences in terms that are both country-specific and universal. Universal teaching competences may be defined as complex combinations of knowledge, skills, understanding, values, and attitudes, leading to effective action, and they include good knowledge of one's subject and curriculum, the ability to plan and teach well-structured lessons, experience in using technology in the classroom, as well as accurate and productive assessment.

Professional competence in educational assessment is one of the universal metrics and teachers worldwide need to demonstrate the knowledge and skills necessary for obtaining information and making educational decisions about students, for evaluating the effectiveness of teaching, and for giving feedback to stakeholders. The range and complexity of assessment procedures, which teachers deal with daily, require the development of literacy in language testing and assessment for effective teaching in the 21st century.

Moreover, teachers are expected to be engaged in the assessment and evaluation of students with special needs, including students with a greater knowledge base and the ability to retrieve information faster and to respond to tasks more efficiently. As such, teachers should be prepared to respond to these new challenges, and this chapter is primarily aimed at defining the concept of linguistic giftedness, describing the ways of identifying linguistically gifted students, developing assessment methods appropriate for them, and scoring and interpreting the assessment results. This chapter also outlines strategies of choosing valid assessment techniques that teachers of English can efficiently use to enhance the unique abilities of these youngsters.

Given the gap between current methodologies and today's changing professional standards, the development of teachers' language assessment literacy has become a challenge in teacher training, in part due to a paucity of course material aimed at developing professional competence in the assessment of gifted students. The focus of the present study is to examine the techniques used to train future teachers of English to design an assessment environment for linguistically gifted students.

2. Defining language assessment literacy

According to the Standards for Teacher Competence in the Educational Assessment of Students (1990), assessment is defined as:

> the process of obtaining information that is used to make educational decisions about students, to give feedback to the student about his or her progress, strengths, and weaknesses, to judge instructional effectiveness and curricular adequacy, and to inform policy.

Assessment literacy is crucial to making decisions about the use of assessment, as well as its development and administration, and for this reason it has become the topic of detailed discussion and extensive research (Stiggins 1991, 1995; Brindley 2001; Hasselgreen, Carlsen and Helness 2004; White 2009; Fulcher 2012; Malone 2013; O'Loughlin 2013; Pill and Harding 2013; Taylor 2013; Tsagari and Vogt 2014, 2017; Harding and Kremmel 2016; Scarino 2017; Hamp-Lyons 2017; Stabler-Havener 2018).

Fulcher (2013) states that the first decade of the 21st century has seen a phenomenal increase of the responsibilities placed upon language teachers in testing and assessment and explains this trend by the growing use and rapid expansion of tests and assessments. Assessment of students is now considered to be an essential part of teaching, as well as a considerable force in fostering and enhancing student learning. Based on his findings, Fulcher (2013) defines assessment literacy (AL) as:

> the knowledge, skills and abilities required to design, develop, maintain or evaluate large-scale standardized and/or classroom based tests, familiarity with test processes, and awareness of principles and concepts that guide and underpin practice, including ethics and codes of practice. The ability to place knowledge, skills, processes, principles and concepts within wider historical, social, political and philosophical frameworks in order to understand why practices have arisen as they have, and to evaluate the role and impact of testing on society, institutions, and individuals.

"Assessment literacy" has become a widely accepted term in educational research, and this has recently extended to the field of language testing through another commonly referenced term, "language assessment literacy" (LAL), which is defined as:

> the ability to design, develop and critically evaluate tests and other assessment procedures, as well as the ability to monitor, evaluate, grade and score assessments on the basis of theoretical knowledge (Vogt and Tsagari 2014) or as a range of skills related to test production, test score interpretation and use, and test evaluation in conjunction with the development of a critical understanding about the roles and functions of assessment within society. (O'Loughlin 2013)

The language assessment community has noted that LAL should be considered separately from general AL (Harding and Kremmel 2016):

> because of the unique complexities that are entailed in the testing and assessment of linguistic skills, knowledge, and communicative competence.

To summarize, language assessment literacy may be understood as indicating a repertoire of competences that enable a teacher to understand, evaluate and, in some cases, create language tests and analyze test data (Pill and Harding 2013), and an assessment-literate person is one who can discern between excellent and poor-quality assessments and apply that knowledge to make informed inferences about student achievement (Stiggins 1991). Moreover, individuals who are literate in language assessment should be able to develop quality instruments of assessment and properly analyze the evidence gathered from the administration of these instruments (Stabler-Havener 2018).

Although there are no teaching knowledge tests or certification programs devoted entirely to measuring teachers' LAL levels, evidence from empirical studies does suggest that there is a lack of sufficient knowledge and skills in matters of assessment, which in turn implies a low level of LAL (Tsagari and Vogt 2017). Training to develop teachers' LAL should become an integral part of pre-service preparation as language teachers will in future be expected to have a range of strategies at their disposal to implement classroom assessment and evaluate its success (Fulcher 2012).

Furthermore, the research of Standards for Teacher Competence in the Educational Assessment of Students (1990) highlights the fact that:

> it is not enough that teachers are able to select and develop good assessment methods; they must also be able to apply them properly. Teachers should be skilled in administering, scoring, and interpreting

> results from diverse assessment methods. They will be able to use accumulated assessment information to organize a sound instructional plan for facilitating students' educational development.

Teachers are expected to exploit the increased variety of assessment procedures, such as the alternative forms of assessment that have become part of "mainstream" practices in many educational contexts (Tsagari and Vogt 2017). Thus, it is crucial for foreign language teachers to be involved in training initiatives to broaden and diversify their assessment literacy with varying priorities depending on contextual assessment requirements (Tsagari and Vogt 2017).

In current language education there is a rich and increasing diversity of educational contexts.

> The implications for working in and with this intensified diversity are challenging for teaching foreign languages, and the implications for assessment are perhaps even more challenging as this context of diversity requires new forms of assessment. (Scarino 2017)

One example of this diversity is the context in which gifted education operates. Many students express a desire for higher achievement and academic success (Stiggins 1995; Rogers 2002), demonstrating special learning needs (Coleman and Cross 2005), which forces teachers to focus on the identification of potentially gifted students and the provision of high-quality educational programs and assessment for nurturing the talents of these students.

3. Describing the construct of giftedness

Research regarding more effective methods for identifying, teaching, and assessing gifted students must entertain the issues related to how giftedness is defined and the attributes used to articulate this definition.

Recent years have witnessed an increase in research on the topic of giftedness and an emergence of a variety of definitions of giftedness (Kitano and Kirby 1986; Frasier, Passow and Garcia 1995; Gallagher 2000; Colangelo and Davis 2002; Castellano and Diaz 2002; Treffinger and Isaksen 2005; Borland 2005; Maker and Schiever 2005; Davis 2006; VanTassel-Baska and Brown 2007; Dai 2010; Renzulli 2012; Ford 2013; George 2013).

> The concept of 'giftedness' refers to the individual potential for high or outstanding achievement in one or more areas of ability (Mönks and Pflüger 2005). Giftedness is also defined as performance that is clearly at

> the upper end of the distribution in a specific talent domain even relative to other high-functioning individuals in that domain. Further, giftedness can be viewed as developmental in that in the beginning stages, potential is the key variable; in later stages, achievement is the measure of giftedness; and in fully developed talents, eminence is the basis on which this label is granted. (Worrell, Olszewski-Kubilius and Subotnik 2012)

Giftedness has many dimensions, including the following: abilities, talent, personality factors, and environment. Identifying, recognizing, positively evaluating, and nurturing giftedness all present an important challenge to educators.

Current trends demonstrate that there is a dynamic growth in the field of gifted education both in Europe and in the USA. The survey, "Gifted Education in 21 European Countries: Inventory and Perspective" (Mönks and Pflüger 2005), reveals a dynamic development of gifted education in European schools:

> Educational legislation in 14 of the 21 European countries in this inventory now refers explicitly to gifted children. Legislation has been introduced in France and Greece that recognises gifted individuals. Positive political attitude towards gifted education came up in Sweden. Provisions for gifted students were improved in the following countries: Austria, Belgium-Flanders, Switzerland, Germany, Denmark, Spain, Ireland, Luxemburg, Romania and the United Kingdom. Substantial improvements in professional training and/or upgrading has been reported in Switzerland, Germany, France, Hungary, Italy (the German speaking area of northern Italy), Luxembourg, Poland, Romania, Sweden and United Kingdom. The legislative status of gifted children and their needs became a reality in these countries.

Similarly, an analysis of the Russian system of education demonstrates that special attention is paid to identifying gifted students and creating conditions for their development, as stated in the special federal program, "Gifted Children" (2003), which is currently being implemented. An expert community of scientists representing the Russian Psychological Society, the Institute of Psychology of the Russian Academy of Sciences, and the Department of Psychology of Moscow State University defined the concept of giftedness as:

> a systemic and developing quality of psyche, allowing a person to achieve significant results in various activities.

They have also documented the fundamental research that has been conducted in the field of giftedness, as well as modern trends in identifying gifted children and developing their talents, based on both positive and negative experiences.

American educational policy is also aimed at recognizing gifted children and providing them with appropriate education services (Robinson, Shore and Enersen 2006). The US Department of Education (1993) defines gifted students as:

> children and youth with outstanding talent who perform or show the potential for performing at remarkably high levels of accomplishment when compared with others of their age, experience, or environment. These children and youth exhibit high performance capability in intellectual, creative, and/or artistic areas, possess an unusual leadership capacity, or excel in specific academic fields. They require services or activities not ordinarily provided by the schools.

According to the results of national research (McClain and Pfeiffer 2012), which examined each of the 50 states' policies specific to the identification of gifted students, 48 states have established definitions of giftedness and 32 states have specific legislative policy mandating the identification of students who are gifted.

The analysis demonstrates that there is a general agreement on the necessity of identifying, recognizing, and nurturing giftedness. In modern psychological and pedagogical studies, various types of giftedness are classified.

4. Conceptualizing linguistic giftedness

Given the concept of giftedness that has been expanded considerably to other domains, the addition of "linguistic" giftedness was to be expected (Winner 1996; Llanes 1979; Bartz 1982; Angelelli, Enright and Valdés 2002; Matthews and Foster 2005; Brulles, Castellano and Laing 2011).

Linguistic giftedness may be understood as a combination of a student's high level of communicative competence in a foreign language, his or her ability to implement cognitive activities by means of a foreign language, and individual psychological characteristics, including creativity in the choice of forms and means of communication. Linguistically gifted students are effective communicators who demonstrate sustainable motivation to study a foreign language.

From the perspective of gifted and talented education, this definition suggests that linguistically gifted students exhibit a range of unique linguistic, memory-analytic, creative-synthetic, and practical-contextual abilities. Such youngsters are able to (Angelelli, Enright and Valdés 2002):

> approach and analyze problems, construct plans, decide which performance components to utilize, select strategies, monitor their behavior and evaluate the processing of information, and keep track of what has been done and what remains to be done in the solution of a problem.

Obviously linguistically gifted students need a stimulating and creative learning environment and assignments designed specifically for their future development. So, an increasing concern nowadays is the education, guidance, and assessment of gifted students who possess a greater knowledge base in foreign languages and are capable of understanding and retrieving information faster, responding to tasks more efficiently, and selecting problem-solving strategies. It is necessary for the teachers who work with these students to be assessment-literate in order to develop efficient and challenging assessment instruments and analyze the assessment results.

The governments of many countries nowadays have established national competitions in foreign languages and linguistic Olympiads for gifted students (UKLO, UK Linguistics Olympiad; Russian National Olympiad in Foreign Languages; NACLO, North American Computational Linguistics Olympiad; OzCLO, Australian Computational and Linguistics Olympiad). These competitions are a proven method of exposing linguistically gifted students to the complexity, beauty and wonder of the world's languages, but they require teachers to develop new assessment skills that they need to enhance learning because, according to Stiggins (1991):

> we must teach teachers and their supervisors to use assessment as a teaching and learning tool–not merely as a grading tool.

The accomplishment of this task is undoubtedly connected with some changes in the professional training and activities of teachers of foreign languages, who educate, support, evaluate, and assess linguistically gifted students. Unfortunately, the analysis performed by Pill and Harding (2013) shows that teachers display either some lack of knowledge or a degree of misconception about language assessment practices and differences between testing instruments. The results of the European survey of language testing and assessment needs (Hasselgreen, Carlsen and Helness 2004) also demonstrate that teachers are in need of training in:

> preparing own tests, assessing aspects of culture, assessing integrated skills, establishing reliability and validity, defining criteria, interpreting results.

Exploring the reasons why many teachers lack the competence to accurately assess their students, it may be stated that language teachers are aware of a variety of assessment needs that are not currently catered for (Fulcher 2012) and they stress that training in language assessment was neglected during their undergraduate studies and pre-service training (Tsagari and Vogt 2017). All sources of data clearly show that teacher education programs do not provide adequate training in language testing and assessment and teachers need to improve their overall competence in language testing and assessment, skills and knowledge.

The paucity of curricula and courses aimed at developing professional competence in the educational assessment of gifted students has motivated this research.

5. Pre-service Teacher Education: A Five-component Professional Development Course

To help meet the demand for the development of teachers' professional competence in the educational assessment of linguistically gifted students a pre-service assessment development course for teachers of foreign languages was designed. This course is an integral part of pre-service teacher education and it consists of five core components (modules) (see Figure 1). When designing this course, an attempt was made to apply a practice-oriented approach to the professional development of teachers of foreign languages and the content of the course enables teachers to meet the needs of linguistically gifted students. The course is also aimed at developing LAL competency among future teachers who might work with linguistically gifted students.

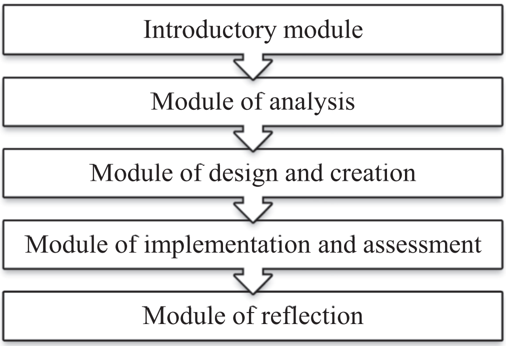


Figure 1: Map of core components of a pre-service assessment development course for teachers of foreign languages

The first introductory module is aimed at familiarizing future teachers of foreign languages with the concepts of assessment literacy and language assessment literacy, giftedness and linguistic giftedness. Moreover, they learn how to classify various intellectual competitions in foreign languages depending on their goals and content. The practical aspects of running competitions in foreign languages are addressed, for example:

- what a competition in foreign languages looks like and what it aims to measure;
- how to find available documentation on language competitions and Olympiads; and
- how the scores are determined and who determines them.

The introductory module is majorly focused on introducing teachers to assessment literacy and developing their understanding of the basic principles of assessment task design.

The module of analysis focuses on two major analytical activities: firstly, trainee teachers learn how to identify linguistically gifted students, determine their individual needs, and having analyzed their written and oral responses choose a strategy to improve the efficiency of teaching for linguistically gifted learners to benefit from it. Secondly, future teachers of foreign languages analyze the tasks designed for various competitions in foreign languages and past papers from previous years. They pay special attention to the format of tasks and forecast the difficulties that gifted students may encounter.

Unfortunately, the majority of teachers are largely unsuccessful in identifying linguistically gifted students and at developing programs that might enhance the unique abilities of these youngsters. Linguistically gifted students distinguish themselves by instrumental and expressive behaviour. They demonstrate the ability to set goals and independently choose activities for mastering a foreign language and developing their subject knowledge; they make full use of their multidisciplinary knowledge and unique problem-solving strategies to ensure that they are able to communicate a message successfully; they are capable of elaborating on ideas following a logical sequence; they mostly reject clichés, stereotypes, and formal answers; and show an inclination for self-study and encourage self-regulation and reflection. Furthermore, linguistically gifted students typically demonstrate the ability to carry out objective self-evaluation and reflection. The expressive behaviour of linguistically gifted students may be characterized by exceptional cognitive needs, curiosity, and increased attention; a high degree of involvement and genuine interest; perfectionism and a strong motivation to study a foreign language.

As mentioned above, linguistically gifted students possess unique skills and abilities, which should be taken into consideration when designing tasks for them. The tasks, therefore, should not only be linguistically advanced, but also assess the intellectual and creative abilities of gifted students, their analytical skills and desire to take the initiative; foresee the high efficiency of problem-solving that these students demonstrate and their capability of understanding and retrieving information faster; and take into account the predominance of cognitive motives, and deep interest in the foreign language and culture. When analyzing past papers, trainee teachers are sure to notice the unusual format of tasks, for example, the use of various integrated activities that differ from the traditional ones and allow students to express their autonomy in the choice of strategies. The brief discussion of skills and abilities of linguistically gifted students shows that they need a special motivating and individualized learning environment which may be created, provided that teachers' professional competence in the assessment of gifted children and their LAL is developed.

The module of analysis is a necessary foundation for the module of design and creation, in which future teachers of foreign languages start designing an educational environment for linguistically gifted students by creating their own tasks and mapping out the strategies to individualize the learning process. At this stage, students select the sources of teaching and reference materials guided by the strategies and principles of selection, decide on the content and develop tasks, which should be aligned with the intended learning outcomes, in the format of intellectual competitions in foreign languages through a case-study approach. In order to do it students need to demonstrate their understanding of learning goals, features of assessment and to apply the principles of assessment task design. As a result, trainee teachers assemble their portfolios of tasks which contain different self-designed tasks and activities, as well as the evidence for reflection, for example: integrated tasks, creative tasks, tasks based on visual forms of presenting information, etc. (See Appendix A.)

The module of implementation and assessment is of particular interest as it is designed to bridge the divide between theory and practice, help nurture trainee teachers' LAL and promote an assessment culture. Future teachers of foreign languages are given the opportunity to apply their knowledge and put assessment into practice during their internships at schools. This hands-on practice is a necessary stage in the development of teachers' assessment literacy as it provides students with an opportunity to introduce their newly designed tasks into classrooms and get feedback. In addition, they take part in assessing the performance of high school students in intellectual academic competitions in foreign languages, and check and grade their papers. Being part of a marking commission, future teachers of foreign languages learn to apply certain criteria when scoring and interpreting the assessment results. Moreover, they analyze the quality of assessment tasks and select students' work samples for the further analysis of possible discrepancies between the scores given by in-service teachers and grading criteria.

The final module of the course, the module of reflection, maps out a follow-up strategy for putting into practice issues raised in the other modules. Trainee teachers analyze the effectiveness of the tasks developed by them and evaluate the results of a student perception survey that they previously conducted in order to collect feedback data. Pre-service teachers are advised to use the Panorama Student Survey developed by the Harvard Graduate School of Education as one of the survey tools. The module of reflection ends with the presentation of the study materials developed by each student and self-evaluation followed by a discussion. If necessary, changes may be made to the developed tasks.

As a result, trainee teachers develop a range of skills related to the choice of valid assessment procedures and techniques, task production, score interpretation, and assessment evaluation. They also gain new experience in assessment practices which leads to a significant improvement in trainee teachers' language assessment literacy.

6. Conclusion

We believe that the training of pre-service teachers constitutes one of the most important aspects in the quality assurance of language assessment and it is essential that gifted education becomes an integral part of the basic curriculum of foreign language teacher training.

The course designed at Herzen State Pedagogical University of Russia was piloted at the Department of Modern Languages with 22 pre-service teachers who practised designing an assessment environment for linguistically gifted students. The pilot course showed that trainee teachers mastered the ability to use different techniques of assessing gifted students, demonstrated the willingness to analyze learning outcomes and use this data when planning their teaching activities. The course also encouraged their creativity in designing tasks. It is assumed that the introduction of similar courses into pre-service teacher education will develop future teachers' LAL competency and facilitate the design and use of high-quality language assessments for linguistically gifted students not only nationally, but internationally too.

To summarize, it is essential for pre-service teachers to understand the concept of the assessment environment for linguistically gifted students and to outline the strategies of choosing valid assessment procedures and techniques, scoring and interpreting the assessment results which may be efficiently used when planning teaching.

Our results indicate that language assessment literacy is one of the fundamental issues in language teacher education nowadays and this course is aimed at developing it and improving the quality of teacher training by preparing pre-service teachers of foreign languages for new challenges in gifted education.

7. References

Angelelli, Claudia, Kerry Enright, and Guadalupe Valdés. 2002. *Developing the Talents and Abilities of Linguistically Gifted Bilingual Students: Guidelines for Developing Curriculum at the High School Level*. University of Connecticut, US: NRC/GT.

Bartz, Walter. 1982. "The Role of Foreign Language Education for the Gifted and Talented Student." *Foreign Language Annals* 15(5):329-334.

Bogoyavlenskaya, Diana B. 2004. "Rabochaya Kontseptsiya odarennosti" [Draft Conception of Giftedness]. *Voprosy obrazovaniya* 2:46-68. Retrieved from High School of Economics website: https://vo.hse.ru/.

Borland, James H. 2003. *Rethinking Gifted Education*. Education and Psychology of the Gifted Series. New York, NY, US: Teachers College Press.

Borland, James H. 2005. "Gifted Education Without Gifted Children: The Case for No Conception of Giftedness." In *Conceptions of Giftedness*, edited by Robert J. Sternberg, and Janet E. Davidson, 1-19. New York, NY, US: Cambridge University Press.

Brindley, Geoff. 2001. "Language Assessment and Professional Development". In *Experimenting with Uncertainty: Essays in Honour of Alan Davies*, edited by Cathie Elder, Annie Brown, Kathryn Hall, Noriko Iwashita, Tom Lumley, Tim McNamara and Kieran O'Loughlin, 126-136. Cambridge: Cambridge University Press.

Brulles, Dina, Jaime A. Castellano, and Peter C. Laing. 2011. "Identifying and Enfranchising Gifted English Language Learners." In *Special Populations in Gifted Education: Understanding our Most Able Students from Diverse Backgrounds*, edited by Jamie A. Castellano, and Andrea D. Frazier, 305-313. Waco, TX, US: Prufrock Press.

Castellano, Jaime A., and Eva Diaz. 2002. *Reaching New Horizons: Gifted* *and Talented Education for Culturally and Linguistically Diverse Students*. Pearson.

Colangelo, Nicholas, ed., and Gary A. Davis. 2002. *Handbook on Gifted Education* (3rd ed.). Pearson.

Coleman, Laurence J., and Tracy L. Cross. 2005. *Being Gifted in School: An Introduction to Development, Guidance, and Teaching* (2nd ed.). Waco, TX, US: Prufrock Press.

Dai, David Y. 2010. *The Nature and Nurture of Giftedness: A New Framework for Understanding Gifted Education.* Education and Psychology of the Gifted Series. New York, NY, US: Teachers College Press.

Davis, Gary A. 2006. *Gifted Children and Gifted Education: A Handbook for Teachers and Parents*. Scottsdale, AZ, US: Great Potential Press, Inc.

Ford, Donna Y. 2013. *Recruiting and Retaining Culturally Different Students in Gifted Education.* Waco, TX, US: Prufrock Press.

Ford, Donna Y., and Michelle F. Trotman. 2001. "Teachers of Gifted Students: Suggested Multicultural Characteristics and Competencies." *Roeper Review* 23(4):235-239.

Frasier, Mary M., Jaime H. Garcia, and A. Harry Passow. 1995. *A Review of Assessment Issues in Gifted Education and Their Implications for Identifying Gifted Minority Students* (Research Monograph 95204). Storrs, CT: The National Research Center on the Gifted and Talented, University of Connecticut.

Fulcher, Glenn. 2012. "Assessment Literacy for the Language Classroom." *Language Assessment Quarterly* 9(2):113-132.

Gallagher, James J. 1991. "Issues in the Education of Gifted Students." In *Handbook of Gifted Education*, edited by Nicholas Colangelo, and Gary A. Davis, 14-23. Boston: Allyn and Bacon.

Gallagher, James J. 2000. "Unthinkable Thoughts: Education of Gifted Students." *Gifted Child Quarterly* 44(1):5-12.

George, David. 2013. *Gifted Education: Identification and Provision*. London: David Fulton Publishers.

Hamp-Lyons, Liz. 2017. "Language Assessment Literacy for Language-learning Oriented Assessment." *Papers in Language Testing and Assessment* 6(1):88-111.

Harding, Luke, and Benjamin Kremmel. 2016. "Teacher Assessment Literacy and Professional Development." In *Handbook of Second Language Assessment. Handbooks of Applied Linguistics*, edited by Dina Tsagari, and Jayanti Banrjee, vol. 12, 413-428. Mouton De Gruyter.

Hasselgreen, Angela, Cecile Carlsen, and Hildegunn Helness. 2004. *European Survey of Language Testing and Assessment Needs. Part One: General Findings.* Retrieved from the website: http://www.ealta.eu.org/resources.htm.

Kitano, Margie K., and Darrell F. Kirby. 1986. *Gifted Education: A Comprehensive View*. Boston: Little, Brown.

Llanes, Jose R. 1979. "Bilingualism and the Gifted Intellect." *Roeper Review* 2(3):11-12.

Maker, C. June, and Shirley W. Schiever. 2005. *Teaching Models in Education of the Gifted* (3rd ed.). PRO-ED.

Malone, Margaret. 2013. "The Essentials of Assessment Literacy: Contrasts between Testers and Users." *Language Testing* 30(3):329-44.

Matthews, Dona J., and Joanne F. Foster. 2005. "Mystery to Mastery: Shifting Paradigms in Gifted Education." *Roeper Review* 28(2):64-69.

McClain, Mary-Catherine, and Steven Pfeiffer. 2012. "Identification of Gifted Students in the United States Today: A Look at State Definitions, Policies, and Practices." *Journal of Applied School Psychology* v 28(1): 59-88.

Mönks, Franz J., and Robin Pflüger. 2005. *Gifted Education in 21 European Countries: Inventory and Perspective*. Radboud Universiteit Nijmegen.

O'Loughlin, Kieran. 2013. "Developing the Assessment Literacy of University Proficiency Test Users." *Language Testing* 30(3):363-380.

Pill, John, and Luke Harding. 2013. "Defining the Language Assessment Literacy Gap: Evidence from a Parliamentary Inquiry." *Language Testing* 30(3):381-402.

Renzulli, Joseph S. 2012. "Reexamining the Role of Gifted Education and Talent Development for the 21st Century: A Four-part Theoretical Approach." *Gifted Child Quarterly* 56(3):150-159.

Robinson, Anne, Bruce M. Shore, and Donna L. Enersen. 2006. *Best Practices in Gifted Education: An Evidence-based Guide*. Waco, TX, US: Prufrock Press.

Rogers, Karen B. 2002. *Re-forming Gifted Education: Matching the Program to the Child*. Scottsdale, AZ, US: Great Potential Press, Inc.

Scarino, Angela. 2017. "Developing Assessment Literacy of Teachers of Languages: A Conceptual and Interpretive Challenge." *Papers in Language Testing and Assessment* 6(1):18-40.

Stabler-Havener, Michelle L. 2018. "Defining, Conceptualizing, Problematizing, and Assessing Language Teacher Assessment Literacy." *Teachers College, Columbia University Working Papers in*

Applied Linguistics and TESOL 18(1):1-22.

Standards for Teacher Competence in Educational Assessment of Students. 1990. Developed by the American Federation of Teachers, National Council on Measurement in Education, National Education Association. Retrieved from the website: https://buros.org.

Stiggins, Richard J. 1991. "Assessment Literacy." *Phi Delta Kappan* 72(7):534-539.

Stiggins, Richard J. 1995. "Assessment Literacy for the 21st Century." *Phi Delta Kappan* 77(3):238-245.

Taylor, Lynda. 2013. "Communicating the Theory, Practice and Principles of Language Testing to Test Stakeholders: Some Reflections." *Language Testing* 30(3):403-412.

Treffinger, Donald J., and Scott G. Isaksen. 2005. "Creative Problem Solving: The History, Development, and Implications for Gifted Education and Talent Development." *Gifted Child Quarterly* 49(4):342-353.

Tsagari, Dina, and Karin Vogt. 2017. "Assessment Literacy of Foreign Language Teachers Around Europe: Research, Challenges and Future Prospects." *Papers in Language Testing and Assessment* 6(1):41-63.

VanTassel-Baska, Joyce, and Elissa F. Brown. 2007. "Toward Best Practice: An Analysis of the Efficacy of Curriculum Models in Gifted Education." *Gifted Child Quarterly* 51(4):342-358.

Vogt, Karin, and Dina Tsagari. 2014. "Assessment Literacy of Foreign Language Teachers: Findings of a European Study." *Language Assessment Quarterly* 11(4):374-402.

White, Eddy. 2009. "Are You Assessment Literate? Some Fundamental Questions Regarding Effective Classroom-based Assessment." *OnCUE Journal* 3(1):3-25.

Winner, Ellen. 1996. *Gifted Children: Myths and Realities.* New York, NY, US: Basic Books.

Worrell, Frank C., Paula Olszewski-Kubilius, and Rena F. Subotnik. 2012. "Important Issues, Some Rhetoric, and a Few Straw Men: A Response to Comments on 'Rethinking Giftedness and Gifted Education.'" *Gifted Child Quarterly* 56:224-231.

Appendix A. Sample Tasks

Integrated listening and reading. You will hear an interview with the gardening experts, Jed and Helena Stone. Listen to the recording[1] and then read a passage on the same topic. For questions 1-4 choose the best answer (A, B, C or D).

A. means that the idea is expressed both in the listening and the reading passage.
B. means that the idea is expressed in the reading passage only.
C. means that the idea is expressed in the listening passage only.
D. means that the idea is expressed neither in the listening nor in the reading passage.

In the UK, there are around 27 million people who partake in gardening. This is a huge portion of the 64 million people that currently reside in this country. Gardening has always been a matter of personal taste, and often the outstanding works of previous generations are torn down to make way for the style of the next. For that reason, it is hard to find unaltered examples of historical gardens in England. Yet, throughout Britain there are gardens great and small, formal and informal, private and public, that illustrate the British passion for creating green, growing spaces of their own. All are different, and all, like their owners and creators, have a distinct personality. The face of British gardening, Jed and Helena Stone, have built the jewel garden over the past decade and it has bloomed from the muddy fields around their farmhouse, a perfect metaphor for their own rise from the ashes of a spectacular commercial failure. Their garden at Longmeadow is laid out in a semi-formal grid pattern and at its heart is what is now known as the jewel garden–a flower garden favouring the rich and vivid colours of jewels. The jewel garden is the example of a creative partnership that has weathered the greatest storm, and a testament to the healing powers of the soil.

1. Both Jed and Helena Stone are proud of the way their public profile reflects their achievements.
2. Gardening is an extremely popular British pastime.
3. The design of the jewel garden was thoroughly planned and meticulously thought through.
4. The name “jewel garden” provided a useful framework for the project.

[1] To play the recording, go to: https://soundcloud.com/perevertkina-m/listening.

Guess the verbs and prepositions in phrasal verbs and then form phrasal nouns from them. There is an example at the beginning (0). Use all of these phrasal nouns (#3, 6, 9, 12, 15) in your short story that you are writing for The Short Story Competition, the contest that spotlights notable new literary talent.

Verb	Preposition	Noun
They never (0) **pass** the news on to me.	Can you stop (0) **by** on your way to school?	A (0) **passer-by** saw the accident and stopped to help.
You agreed to help me and you can't (1) _____ out now!	I tried to call him but he didn't pick (2) _____.	Before we turn off our computer, we will make a (3) _____ of all the files.
On Sundays I stay in bed for as long as I want and (4) _____ up at midday.	He is not going to give (5) _____ any plot details. Let's ask somebody else.	Millions of holidaymakers are heading off by road today for a Christmas (6) _____.
I'm worried that our fridge might (7) _____ down after the guarantee has expired.	While I was waiting for you, I've looked (8) _____ some magazines.	Their research may lead to an important technological (9) _____.
Jobs and Wozniak (10) _____ up a shop in Jobs' garage, and began working on the Apple I.	It was cold outside and she had to put (11) _____ a coat.	The sudden (12) _____ of autumn has taken tens of thousands by surprise.
Children are asked to (13) _____ up when the teacher comes into the classroom.	Let's meet at your place for coffee. I'll come (14) _____ on Friday.	If there is a (15) _____, get him or her to phone for an ambulance at once.

Look at the map of London. Identify and label the famous sights. Then read the descriptions of some of these sights and fill in the gaps with one or two proper nouns. The number of letters is given and there is an example at the beginning (0). You will need this information to make a presentation. Imagine you are a tour guide taking your partner on an excursion to the centre of London. Give a tour and then answer your partner's questions.

Built to commemorate Lord (0) **N e l s o n**’s victory over the (1)_ _ _ _ _ _ and Spanish at Trafalgar in 1805, Trafalgar Square is one of London’s most popular tourist spots. Notable monuments in the square include statues of Henry Havelock, General Gordon, Charles James Napier and an equestrian statue of (2)_ _ _ _ _ _ IV. The biggest attraction, however, is (0) **N e l s o n**’s Column. Constructed entirely from granite, this 185 ft high monument to (3)_ _ _ _ _ _ _’s greatest hero overlooks the square’s fountains and bronze reliefs, cast from French cannons and depicting (0) **N e l s o n**’s victories at Cape St Vincent, the Nile, Copenhagen and Trafalgar. Buildings surrounding Trafalgar Square include the imposing Admiralty Arch.

The (4)_ _ _ _ _ _ _ _ _ _ Bridge connects two tourist areas across the river (5)_ _ _ _ _ _. The southern end of the gently swooping suspension bridge is located near the new (6)_ _ _ _ _Theatre and the (7)_ _ _ _ _ _ _ _ _ _ _ Gallery. The northern end sits near the imposing (8)_ _ _ _ _ _ _ _ _ _'s Cathedral. Pedestrians can gain a wonderful view of the cathedral's dome from the bridge and the sight is especially marvellous at night.

The (9)_ _ _ _ _ Bridge, (10)_ _ _ _ _ _’s most famous bridge, and the Shard, the city's tallest building, are also clearly visible from here.

Read the sonnet below and define the highlighted words from the sonnet. You will find all the necessary answers in the word cloud (tree). There is an example at the beginning (0).

Shall I compare thee to a summer's day?
Thou art more lovely and more temperate.
Rough winds do shake the darling buds of May,
And summer's lease hath all too short a date.

Sometime too hot the eye of heaven shines,
And often is his gold complexion dimmed,
And every fair from fair sometime declines,
By chance, or nature's changing course untrimmed.
But thy eternal summer shall not fade,
Nor lose possession of that fair thou ow'st,
Nor shall death brag thou wander'st in his shade,
When in eternal lines to time thou grow'st,
So long as men can breathe, or eyes can see,
So long lives this, and this gives life to thee.

(0)	darling	=	pretty
1	temperate	=	
2	the eye of heaven	=	
3	fair	=	
4	ow'st	=	

Now decide whether the sentences below are true or false according to the sonnet you've read. There is an example at the beginning (0).

(0)	Aging brings natural changes.	**TRUE**	FALSE
1	May was a summer month in Shakespeare's time.	TRUE	FALSE
2	The youth's beauty is less perfect than the beauty of a summer day.	TRUE	FALSE
3	The poet is full of confidence that his poetry will live as long as there are people on the Earth.	TRUE	FALSE

Pay attention to the personal pronouns in the sonnet. Find the necessary pronouns and fill in the gaps in the table below. There is an example at the beginning (0).

Person	Number	Subject	Object	Possession
1st person	Singular	(0) **I__**	Me	My
2nd person	Singular	(1) _____	(2) _____	(3) _____
3rd person	Singular	He/She/It	Him/Her/It	His/Her/Its

CHAPTER THIRTEEN

THE WASHBACK EFFECT AND TEACHERS' ASSESSMENT LITERACY: THE STEPPING STONE FOR MORE LEARNER-CENTRED EXAM-CLASSES

IRINI-RENIKA PAPAKAMMENOU

Exams and especially high-stakes exams exert a powerful influence on both the teaching and learning processes. Teachers are at the grass roots of experiencing test washback at the classroom level and for that reason language assessment literacy should incorporate washback. Washback research has a lot to offer teachers who prepare students for high-stakes exams since they are the most important influential agents in the process. Thus, language assessment literacy is critical to familiarize teachers with the influence of exams and provide practical knowledge so that teachers promote positive washback. This chapter aims to highlight the importance of the washback effect in exam preparation classes and define its scope and nature to teachers. Also, it will identify areas in which washback has been noted by studies relating to what and how teachers teach. It will help teachers understand how washback can provide them with a clearer idea of their role among students and parents and the decisions they can make concerning preparation classes. It will focus on the influence of exams on teaching practices (activities, tasks and teaching strategies). In addition, it will present various factors that contribute to how teachers teach exam preparation classes. It will show the way and the degree that these factors influence the teaching process in order to help them to understand how best to facilitate learning. Suggestions on how to decrease or even eliminate negative washback as well as ways to employ communicative methodologies and alternative assessment techniques will be provided. Finally, this chapter will provide a useful guide to teachers who prepare students for exams since it will include step-by-step guidance on how to "use" the washback effect to create more learner-centred exam classes.

1. Introduction

Language testing plays a very important role in language teaching since testing is crucial to identify the effectiveness of both teaching and learning and thus can be regarded as inseparable from them. As Bachman and Palmer (1996) point out, language testing can provide valuable information about teachers' instruction and language learning. Therefore, a number of stakeholders are affected such as students, testers, institutions, policy makers, curriculum designers, material designers, administrators and, above all, teachers. Basically, as Cheng (2014) supports, the consequences of testing on teaching and learning should be understood and evaluated by all testing stakeholders and all the stakeholders including teachers, students, parents and test makers should collaborate to achieve this (Shohamy 2001).

In this framework, washback research, which explores the potential effects of high-stakes tests on language teaching and learning, constitutes a significant factor in the language testing context. In particular, it has been observed that washback has great impact on exam preparation classes and this impact can have many dimensions (Wall and Alderson 1993). There is significant differentiation of the washback effect on teaching practices compared to content, skills and materials and this is attributed to a number of factors, which are the focus of the present study. Altogether, the complexity of washback is highlighted by the various models and research findings and it is evident in the interactions between societal and school factors (Alderson and Wall 1993; Green 2006; Glover 2006; Tsagari 2009). Washback research also indicates that "teachers play an important role in fostering different types of washback and […] an influential role when test changes are introduced" (Papakammenou 2008b, 330). Language and testing institutions need to make teachers aware of their impact on bringing changes to the classroom or even to an institution (Djurić 2008). Language assessment literacy (LAL) is critical in familiarizing teachers with the necessary changes, which are the results of washback, and thus improve teaching and testing processes and also introduce changes in the educational policy.

2. Washback and LAL

Washback is "the influence of testing on teaching and learning" (Alderson and Wall 1993, 115). Wall (1997) and McNamara (2004) addressed the test impact and the effects of testing beyond the classroom settings and, more specifically, the effects on individuals, policies or practices within, a school, the educational system or society as a whole. Washback can be strong or weak, depending on the effect that a test might have. If a test has a strong effect then it "will determine everything that happens in the classroom and lead all teachers to teach in the same way toward the exam" (Watanabe 2008, 20). *Intensity* (Cheng 2005) and *direction* (Alderson and Wall 1993) are two other elements of washback. Of great importance though is that washback can be both positive and negative. Positive washback effects mean that tests beneficially influence learners, teachers and the teaching context (Alderson & Wall 1993). Harmful or negative washback of tests on teaching and learning is when the test fails to reflect the learning principles (Cheng 2005).

The findings of washback studies have formulated models which highlight areas that can be influenced by tests and factors that contribute to this influence. Each model adds information to the existing ones and thus helps researchers observe areas where washback exists and teachers aim at more positive washback. Alderson and Wall (1993) formulated their hypothesis, emphasizing different factors such as behaviours, attitudes, test consequences and the diverse effects on different stakeholders. This hypothesis was refined by Alderson and Hamp-Lyons in 1996 adding that the amount and type of washback may differ among teachers and learners. Hughes (1993) in the meantime introduced a tripartite model for studying washback and made a distinction between participants, processes and products. Bailey (1996) presented a more complex model of washback adding a variety of sub-categories to Hughes' three components and focusing on the interaction among the aforementioned components, which emphasizes the complex nature of washback. Tsagari (2009) proposed a model of washback which not only addresses the complex nature of washback but also elaborated on the sources of this complexity that derive from both within and outside the classroom context. Green's (2006) model added another dimension to participants, which incorporates the stakes of the exams. Cheng's (2005) model connected both teachers and students' classroom behaviour to the curriculum as planned according to the exams. Finally, according to Glover's (2006) model, the pedagogical skills and knowledge, instead of the examination, influenced the teacher's pedagogical skills.

Studies on washback are significant for language assessment literacy since they can provide valuable information about exam preparation classes, the nature of washback in such classes, teachers' weaknesses in dealing with such classes or coping with negative washback or changing the effect, and teachers' lack of knowledge of assessment practices. In providing evidence of washback, teachers realize the negative impact of testing on their methodologies and of these on learning. If teachers can realize the effects of tests on their teaching and how valid the tests are, they will be able to understand if it is the tests or other factors that influence the nature of exam preparation classes. That way they will make better use of language assessment literacy programs and they will be able to accommodate these programs to their needs.

One of the most significant negative effects of washback is that of constraining the curriculum. Studies (Wall and Alderson 1993; Watanabe 1996, 1997, 2000; Alderson and Wall 1993; Green 2006; Hayes and Read 2004; Tsagari 2011) have shown that the content of exam preparation classes resembles the exams since teachers limit the teaching of skills to those tested in the exams. Another negative influence of exams is on materials (Watanabe 2000; Hawkey 2006; Al-Jamal and Ghadi 2008; Tsagari 2009, 2011). Materials appear not only to be affected by the exams since they reflect the exams, but also to influence the structure of exam preparation classes and more specifically the teaching content since there has been found to be a heavy use of exam-related materials. A test design can also determine to a great extent the skills that students cultivate in class. Washback research has identified an increased focus on skills that are tested in the exams (Aftab et al. 2014; Akpinar and Cakildereb 2013; Mickan and Motteram 2009) and the neglect of skills which are not tested in the exams (Nikolov 1999). Not only teachers but also other stakeholders benefit from these studies, the results of which highlight the negative impact of tests and how tests can alienate test designers, material developers and test takers from their learning principles and educational goals. On the other hand, washback studies that have revealed positive washback have helped teachers to develop specific skills and employ certain methodologies that promote success and enhance language learning. Washback studies can provide information for the quality of a test which if it is "a good test" can then bring beneficial changes to teaching and teachers' instruction (Hughes 1989; Bailey 1996; Wall 2000). Beneficial washback can be achieved by being aware of the factors contributing to a positive washback and taking advantage of them.

As has been mentioned in the previous section, a huge contribution of washback studies is to the assessment that teachers use in exam preparation classes and to language assessment literacy. Conducting research on assessment provides insights into how teachers assess their students. Classroom assessment is one of the most essential and popular teaching strategies that teachers use in classes and an important factor which influences teachers' teaching and students' learning (Cheng et al. 2008). Research has shown that teachers lack essential assessment skills while administrators have low levels of assessment literacy (Vogt and Tsagari 2014). Teachers, regardless of their experience and their studies, do not possess adequate assessment knowledge (Crusan et al., 2016; Tsagari and Vogt 2017). According to Stiggins (2001), assessment illiteracy not only gives false results to students but also prevents students from reaching their full potential. Unfortunately, language programs do not meet the needs with respect to language assessment literacy (Zhang and Yan 2018). LAL is essential not only for language teachers who are the direct test users (Scarino 2013) but also for other stakeholders in helping them understand the scope of this field (Taylor, 2009).

Findings from studies have shown that the assessments that teachers conduct are in close alignment with the final standardized examination and result in the use of practice tests or past papers that resemble the exam (Kiomrs 2011, Papakammenou 2016). Teachers also do not seem capable of preparing tests themselves or critically evaluating their tests or even the standardized tests (Vogt and Tsagari 2014). Most teachers do not feel adequately prepared for their everyday assessment practices. They usually have only a vague idea about assessment (Vogt and Tsagari 2014) which is reflected in their assessment practices, even though they may be aware of the vital roles of reliability, validity, fairness and authenticity to assessment (Jannati 2015). Also, teachers favour summative assessment and use it in a summative way instead of a diagnostic tool (Hidri, 2016).

Research so far has also highlighted the need for basic and advanced training on assessment. Teachers express a need for basic or more advanced training in applying assessment concepts and techniques, as well as making assessment-related decisions (Mertler and Campbell 2005; Vogt and Tsagari 2014; Kiomrs 2011). There is also a need for practical work not only on skills or micro-linguistic aspects (Hasselgreen et al. 2004) but also on more innovative areas such as alternative assessment techniques (portfolios, self- and peer-assessment) (Vogt and Tsagari 2014). Vogt et al. (2018) also report that teachers do not feel confident in areas such as alternative assessment (self-assessment, peer assessment, portfolio assessment), the CEFR and assessment for specific learning difficulties (SpLDs) and thus training is more demanding in these areas. Training is also needed on feedback since teachers lack the ability to communicate test results and measure assessment procedures (Vogt and Tsagari 2014). Regarding measurement concepts, Zhang and Yan (2018) point out that teachers may benefit more from principles of assessment practices rather than theoretical ones and that is why programs should promote the experiential approach to assessment literacy training. Washback studies have shown that there is a range and number of stakeholders who require a level of assessment literacy. Not only those who are professionally involved with testing and assessment and teachers, but also university admissions officers, policy makers and government departments are in need of assessment literacy (Taylor 2009).

To sum up, washback studies focus on the direct washback of standardized assessment frameworks and provide useful findings on teachers' instruction, assessment and needs which have formulated most of the language assessment literacy (LTA) theory. Thus, more investigation into washback is needed on elements that can enhance teachers' assessment training. This study is aimed at distinguishing the types of teaching practices employed by teachers and identifying the factors that contribute to or inhibit the intended washback of multi-exam classes making considerable contributions to different aspects of washback and LTA.

3. The current study

The present study is aimed at examining the washback effect of multi-exam classes and comparing them with one-exam classes. The focus is on the factors that influence the instruction of English as a foreign language in exam preparation classes. In order to achieve this, an understanding of how teachers teach precedes the study dividing teaching practices into "teaching strategies", "activities", and "tasks" (Papakammenou 2018b) and thus providing a complete picture of the components of teachers' instruction.

4. Method

This study uses a mixed-method research design to investigate the washback phenomenon in the Greek EFL multi-exam preparation context. The qualitative data were collected through a case study methodology (Wallace 2000; Yin 2003; Dörnyei 2007) to investigate the nature of the washback effect. The quantitative data were collected through a questionnaire, the questions of which were formulated from the findings of the qualitative data. The findings of the case study will be analyzed and discussed here since they assist in discussing the connection of washback and assessment literacy.

Setting

This study was carried out at a foreign language centre in Greece. Such language schools are private and called *frontistiria*. The specific language centre offered not only general English classes but also prepared teenage and adult students for EFL language certificates. In fact, the end of the general English language lessons and the accomplishment of such classes are signified by success in high-stakes exams. Classes in *frontistiria* take place after school hours mainly in the afternoons and they last nine to ten months. So, in the specific *frontistiria* both teenage and adult classes were investigated at the same level, and the classes were taking place in the afternoon after the completion of the public school program, which played a significant role as the findings show.

B2-level and C2-level language examinations are the two most popular exam levels in Greece and *frontistiria* dedicate a whole year for the preparation of students for these two exam levels. In the specific study at the B2 level, as described within the CEFR (independent users' vantage level), the exam preparation class was investigated since it is the minimum goal for children of the average Greek family (Dendrinos et al. 2013, 16) and thus the most popular one. There are a number of language certificates at both levels that are recognized by the Greek state and accepted as a language qualification in the public sector. There are 24 B2-level language exams that are recognized by the Greek state and even though they are different in content, format and marking criteria, they are all considered to be equal when used in the public sector as formal qualifications.

This number of English language exam certificates has created an interesting educational context which seems to be promising since more and more language examinations enter the market globally. Apart from the internationally recognized English language examinations which are delivered by Cambridge and ETS, for example, washback studies have shown the existence of other high-stakes exams which are used in different national and educational contexts to certify English language knowledge (Alderson and Wall 1993; Alderson and Hamp-Lyons 1996; Read and Hayes 2003; Green 2006; Mickan and Motteram 2008; Tsagari 2009). In the specific setting, 21 language certificates were recognized and therefore students had 21 English language certificates to choose from. Some schools limit the choices to students and others allow them to choose any of the 21 available. Therefore, multi-exam classes are formulated to make a rather interesting educational setting. Students of such classes can choose to sit for any of these certificates depending on their strengths and weaknesses and teachers are obliged to teach any of the certificates that their students have chosen in the same class. This means that teachers may prepare students for two, three or even four exams in the same class. Teachers in the specific *frontistirio* prepared students for three exams, the University of Michigan (ECCE), Pearson's PTE General Exam Level 3 and the Test of Interactive English (TIE). Not only are the format and marking criteria of these exams different but they also test different skills, for example the TIE examination tests only writing and speaking skills. However, this *frontistirio* offered one exam class, too. This class consisted of adult students as opposed to the multi-exam class which consisted of teenagers. The teachers participating in the study taught both multi-exam and one-exam classes and therefore comparisons could be made.

4.1 Participants

The teachers selected to take part in this study were two teachers who taught both multi-exam and one-exam classes as well as other non-exam classes. In this way, comparisons (Yin 2010) between the classes could be made, all exams, taught separately or in combination with others, could be investigated and teachers' teaching methodology could be compared. The first teacher was the principal of the *frontistirio* with more than 40 years of teaching experience. She was Greek-American and had a degree in English literature. The second teacher was also Greek-American and she had a C2-level English language certificate from the University of Michigan. She had taught English for more than 17 years.

4.2 Instruments and Procedure

A variety of data collection methods (Yin 2003) appropriate for washback studies (Shih 2010; Johnson et al. 2005; Yin 2010, Saville 2012) were used. The exam preparation year was investigated as a whole. Interviews with the teachers were conducted at the beginning and the end of the school year (Papakammenou 2016). Time was spent with teachers before the study, in lesson breaks and in meetings for the specific exam classes. Such socialization was significant to create relationships and a feeling of trust with the teachers and gather more information about issues, such as exam decision-making, choice of coursebooks and any difficulties they faced.

A survey instrument was developed for the classroom observations to address the research questions: "What kinds of teaching practices do teachers use in multi-exam classes? Are there any differences between the 1st and 2nd terms?" and "What is the nature and scope of the washback effect on teachers' teaching practices in multi-exam classes?" More specifically, the tasks, activities and teaching practices that teachers used in exam-preparation classes and the nature, scope and degree of the washback effect on teachers and the exam-preparation classes were recorded. The piloted observation scheme (Papakammenou 2016) was adapted from the Communicative Orientation of Language Teaching (COLT) observation scheme (Spada and Fröhlich 1995). The A3 observation scheme consisted of three main parts, each one dedicated to tasks, activities and teaching strategies respectively. A video recorder was used with the consent of the teachers.

Follow-up interviews were conducted immediately after each observation in order to keep the quality of the retrospective data high (Mackey and Gass 2005; Dörnyei 2005). These stimulated recalls were video recorded to capture the teachers' reactions and record their views on the lessons and their methodology. The observation scheme was also used to help teachers with recall. These interviews were used to discover how teachers decide on the teaching practices they use in multi-exam classes. The follow-up interviews were aimed at gaining more accurate descriptive data from the teachers. These qualitative data constitute the backbone of the study giving a better and clearer description, as well as an explanation of the washback effect. The techniques of thematic categorization from the interviews, observations and follow-up interviews were employed using Atlas.ti (Muhr and Freise 2004) for coding and nodding the qualitative data.

As already mentioned, the researcher used observational methods to collect ground notes on teachers' methodology. The classes lasted nine months and were divided into three semesters. The first was a general preparation for B2-level exams using a general B2-level coursebook, the second was the actual preparation for specific exams using practice tests and past papers and the third was the last month in which students should sit the exams. 24 lessons were observed in both the first and second semesters, 12 for each type of class (multi-exam and one-exam classes). In total, 48 lessons were observed and 48 stimulated recalls were conducted.

5. Results and Discussion

The teacher is an important agent and, in some cases, the ultimate one when it comes to classroom-based LTA (Vogt and Tsagari 2014). The findings of the studies showed that there are various factors that influence teachers directly or indirectly (Papakammenou 2018b), which, if treated carefully, can bring more positive washback. In fact, these factors were found to be interrelated and can be classified into six main categories: students, teachers, language learning, school-class/social, materials and the exam itself. Teacher-direct factors are the teachers, students and language learning, because teachers first focus on them and can exert control over them. Thus, teacher-indirect factors such as the school, class, exams, and materials influence teachers indirectly and are controlled by others such as examination boards and publishers. The exam, however, can be considered a category by itself since its role is significant on its own. The fact that tasks and activities were influenced more by indirect factors such as the class, materials and the exam and less by direct factors, and that teaching strategies were affected more by teacher-direct factors such as teachers, students and language learning may indicate significant areas which can bring change to teachers' instruction and, by extension, to LAL.

If LAL training programs provided teachers with practical knowledge on how to change this influence, then a more learner-centric approach to learning could be achieved. By providing teachers with a range of tasks and activities and with ways that these can be used in different stages of the lesson, teachers will be able to use tasks and activities to enhance learning rather than serve the exam needs. Another suggestion is the combination of tasks and teaching strategies so that they serve each other and therefore have been influenced by all groups of factors instead of only one. By proposing teaching strategies that can involve learners in the teaching process by personalizing the materials, including moments of fun, and selecting and evaluating the materials, the exam influence can be eliminated, and students can be placed at the centre of the teaching process.

The findings related to teachers' assessment and feedback allow improvements on assessment to be made as well. The heavy use of assessment materials resembling the exams and the one-sided feedback coming from teachers only indicate the need for more student-centred assessments. Classroom-based assessment training should focus on new and alternative forms of assessment practices. Formative assessment fosters motivation and promotes an understanding of goals and criteria, dynamic assessment foregrounds future development since "it provides mediation that is constantly adjusted and attuned to the learner's responsiveness" (Poehner and Lantolf 2005, 252), diagnostic assessment can help identify students' needs since it pinpoints strengths and weaknesses which can help teachers to implement the necessary changes and other alternative techniques, such as self- and peer-assessment, and portfolios can involve students in the assessment process and enhance learner autonomy. Student involvement in the creation of assessment materials and criteria by providing personal information and experience and thus making them more personal and direct to their needs, life and interests can be a very promising change. This need for language assessment literacy further corroborates the findings of Vogt and Tsagari's (2014) and Khadijeh and Amir's (2015) studies.

This research on washback has shown that the washback effect on skills taught in exam preparation classes varies between teachers, the format of exams, the number of exams and time–whether it is close to the exams or not (Papakammenou 2018a). In multi-exam classes, all skills are taught and a variety of activities and tasks is available. This allows more opportunities for language learning whereas the content and teaching practices used in one-exam classes resemble the exam, and the teaching is routinized in both semesters. Training in the four skills and specifically in the integration of these skills so that teachers will be better equipped with the confidence to apply them in class is also necessary. Teachers should be informed about the positive washback that multi-exam classes can produce since students get involved with various tasks and activities and all skills are taught. Teachers exhibit positive attitudes towards multi-exam classes as they believe that the multi-exam teaching context offers students a better opportunity to experience a more effective and holistic learning process than one-exam classes. They feel that taking advantage of the best features of each exam can actually advance their students' ability to use the language. On top of that preparation for multiple exams it might even be seen to revolutionize the teaching context, making it more "multi-cultural" as students are simultaneously exposed to different varieties of English and different Anglophone cultures on an equal footing. LAL programs could suggest the creation of multi-exam classes in exam preparation courses in order to eliminate negative washback. Such classes can also benefit multicultural and intercultural courses. By allowing students to have more exam choices such classes can become more beneficial for students. Test constructors should also aim at different test formats which will serve test takers' needs and the future use of tests.

However, the findings indicated that multi-exam classes can create negative washback on the teachers' beliefs. Such negative washback is predominately associated with the additional workload entailed by the multi-exam class. Participant teachers stressed the level of exhaustion which stems from the need to organize such classes more carefully because of time constraints and further pressure on both teachers and students to choose the appropriate exam. LAL should contain consultations which will equip teachers with the necessary knowledge on how to organize multi-exam classes, how to choose exams and how to make use of this context to create more learning opportunities for students.

Of particular importance is teachers' reaction to this study. Comments from the data to illustrate this point are: "you made me think of how I can make this class better" and "I did not expect to teach this way, now I know what I want to change". These comments are associated with the importance of washback studies. Washback research should not only be delivered by researchers but also by teachers. LAL can assist teachers to understand the principles of the washback effect and urge teachers towards a self-reflection of investigating washback combining teacher research, which refers to systematic inquiry by professional teachers in any discipline who investigate their own practices (Borg 2013), with washback research. In this way teachers will be able to adapt their teaching practices and make the necessary changes which will target positive washback.

6. Conclusion

As Djuric (2008) points out washback encourages change and can achieve efficient and productive communication between teachers and testers. The present study has accomplished this by providing appropriate knowledge of significant areas concerning washback to equip teachers with tools to hinder or eliminate the negative washback effect on their teaching and students' learning.

LAL is of crucial importance to both in-service and pre-service teachers since if teachers are assessment literate, they will succeed in increasing the positive washback of tests. Practical workshops and theoretical assessment courses that address teachers' immediate needs, such as coping with multi-exam classes, can improve exam preparation courses and teachers' instruction in such classes. Emphasizing a variety of tasks and activities, detailed score reporting and meaningful feedback, alternative assessment practices, an understanding of the testing criteria by the teachers and students, the design of tests and other assessment practices and the improvement of pedagogical skills can help teachers support and improve students' learning and create more student-centric exam classes.

The washback effect has proved to be an important element of LAL and should be incorporated in LAL courses. It should be promoted to teachers as a self-reflection and evaluation mechanism for change and development. Also, washback research should be encouraged to provide more information on a variety of contexts that may exist all around the world. Future research could further work on a wider basis of exams, gathering evidence from other high-stakes tests with different formats. Not only different high-stakes tests but also different cultures can increase the generalizability of the washback effect of public examinations on language education since it allows for replication. Also, the use of activities and tasks that promote language learning and student participation should be investigated and their effectiveness on exam success should be measured. Researchers should try to apply new and alternative types of tasks and activities in exam preparation classes and study their effectiveness on students, language learning and exam success. Finally, students' involvement should be investigated not only in assessment practices and the materials used but also in the decisions of exam selection and the organization of exam preparation classes. By engaging students in teaching practices, by allowing them to personalize materials and set up their own learning, the course's learning objectives can more successfully be achieved.

7. References

Aftab, Asma, Qureshi Sabeen, and William Isabela. 2014. "Investigating the Washback Effect of the Pakistani Intermediate English Examination." *International Journal of English and Literature* 5(7):149-154.

Akpinar, Kadriye Dilek, and Bekir Cakildereb. 2013. "Washback Effects of High-stakes Language Tests of Turkey (KPDS and ÜDS) on Productive and Receptive Skills of Academic Personnel." *Journal of Language and Linguistic Studies* 9(2):81-94.

Alderson, Charles J., and Diane Wall. 1993. "Does Washback Exist?" *Applied Linguistics* 14(2):115-129.

Alderson, Charles, and LizHapm-Lyons. 1996. "TOEFL Preparation Courses: A Study of Washback." *Language Testing* 13:280-297.

Alderson, Charles, and Diane Wall. 1993. "Does Washback Exist?" *Applied Linguistics* 14:115-129.

Alderson, Charles J., and Liz Hapm-Lyons. 1996. "TOEFL Preparation Courses: A Study of Washback." *Language Testing* 13(3):280-297.

Al-Jamal, Dina, and Ghadi Nedal. 2008. "English Language General Secondary Certificate Examination Washback in Jordan." *The Asian EFL Journal* 10(3):158-186.

Bachman, Lyle, and Adrian Palmer. 1996. *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford: Cambridge University Press.

Bailey, Kathleen M. 1996. "Working for Washback: A Review of the Washback Concept in Language Testing." *Language Testing* 13:257-279.

Borg, Simon. 2013. *Teacher Research in Language Teaching: A Critical Analysis*: Cambridge, UK: Cambridge University Press.

Cheng, Liying, Todd W. Rodgers, and Xiaoying Wang. 2008. "Assessment Purposes and Procedures in ESL/EFL Classrooms." *Assessment and Evaluation in Higher Education* 33(1):9-32.

Cheng, Liying. 2005. *Changing Language Teaching Through Language Testing: A Washback Study. Studies in Language Testing 21*. Cambridge: Cambridge University Press.

Cruscan, Deborah, Plakans, Lia, & Gebril, Atta. 2016. "Writing Assessment Literacy: Surveying Second Language Teachers' Knowledge, Beliefs, and Practices." *Assessing Writing* 28:43-56.

Dendrinos, Bessie, Keti Zouganelli and Evdokia Karavas. 2013. *European Survey of Language Competences: Greek National Report.* Athens: University of Athens, RcEL Publications.

Djurić, Melita. 2008. "Dealing with Situations of Positive and Negative Washback." *Scripta Manent* 4(1):14-27.

Dörnyei, Zoltán. 2005. *The Psychology of the Language Learner: Individual Differences in Second Language Acquisition*. Mahwah, NJ: Lawrence Erlbaum.

—. 2007. *Research Methods in Applied Linguistics*. Oxford: Oxford University Press.

Glover, Philip. 2006. "Examination Influence on How Teachers Teach: A Study of the Teacher." PhD dissertation. University of Lancaster.

Green, Anthony. 2006. "Washback to the Learner: Learner and Teacher Perspectives on IELTS Preparation Course Expectations and Outcomes". *Assessing Writing* 11:113-114.

Hasselgreen, Angela, Celile Carlsen, and Hildegunn Helness. 2004. "European Survey of Language Testing and Assessment Needs. Part One: General findings." Accessed March 10, 2019. http://www.ealta.eu.org/resources.htm.

Hawkey, Roger. 2006. *Impact Theory and Practice: Studies of the IELTS Test and Progetto Lingue 2000*. Cambridge: CUP.

Hayes, Belinda, and John Read. 2004. "IELTS Test Preparation in New Zealand: Preparing Students for the IELTS Academic Module." In *Washback in Language Testing; Research Context and Methods*, edited by Liying Cheng, Yoshinori Watanabe, and Andy Curtis, 19-36. New Jersey: Lawrence Erlbaum Associates, Inc.

Hidri, Sahbi. 2015. "Conceptions of Assessment: Investigating what Assessment Means to Secondary and University Teachers." *Arab Journal of Applied Linguistics* 1:19-43.

Hughes, Arthur. 1989. *Testing for Language Teachers*. Cambridge: Cambridge University Press. Hughes, Arthur. 1993. "Backwash and TOEFL 2000." Unpublished manuscript, University of Reading.

Jannati, Seyran. 2015. "ELT Teachers' Language Assessment Literacy: Perceptions and Practices." IJRTE. Accessed March 10, 2019.http://dergipark.ulakbim.gov.tr/ijrte/article/view/5000106872/0.

Johnson, Karen E., Stefanie Rehn Jordan, and Matthew E. Poehner. 2005. "The TOEFL Trump Card: An Investigation of Test Impact in an ESL Classroom." *Critical Inquiry in Language Studies* 2:71-94.

Khadijeh, Bayat, and Rezaei Amir. 2015. "Importance of Teachers' Assessment Literacy." *International Journal of English Language Education* 3(1):139-146.

Kiomrs, Razavipour, Riazi Abdolmehdi, and Naser Rashidi. 2011. "On the Interaction of Test Washback and Teacher Assessment Literacy: The Case of Iranian EFL Secondary School Teachers." *English Language Teaching* 4(1):156-161. Accessed March 10, 2019. http://www.ccsenet.org/journal/index.php/elt/article/view/9677.

Mackey, Alison, and Susan Gass. 2005. *Second Language Research: Methodology and Design.* Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

McNamara, Tim. 2000. *Language Testing*. Oxford: OUP.

Mertler, Craig, and Cynthia Campbell. 2005. "Teachers' Measuring and Knowledge Application of Classroom Assessment Concepts: The Development of Assessment Literacy Inventory." Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada. April 11-15.

Mickan, Peter, and Johanna Motteram. 2004. "An Ethnographic Study of Classroom Instruction in an IELTS Preparation Program." *IELTS Research Reports* 8:17-43.

Muhr, Thomas, and Susanne Freise. 2004. *User's Manual for Atlas.ti 5.0* (2nd ed.). Germany: Scientific Software Development.

Nikolov, Marianne. 1999. "Classroom Observation Project." In *English Language Education in Hungary: A Baseline Study*, edited by Hajnal Fekete, Éva Major, and Nikolov Marianne, 221-246. Budapest: British Council.

Papakammenou, Irini. 2016. "A Washback Study of the Teaching Practices Used in EFL Exam Preparation Classes in Greece." In

Classroom-based Assessment in L2 Contexts, edited by Dina Tsagari, 118-137. Newcastle-upon-Tyne: Cambridge Scholars Publishing.

—. 2018a. "Washback on Language Skills: A Study of EFL Multi-exam Preparation Classes." In *Revisiting the Assessment of Language Abilities,* edited by Sahbi Hidri, 437-454. Springer International Publishing.

—. 2018b. "Examining Washback in EFL Multi-exam Preparation Classes in Greece. A Focus on Teachers' Teaching Practices." In *Teacher Involvement in High Stakes Language Testing*, edited by Daniel Xerri, and Patricia Vella Briffa, 321-339. Springer International Publishing.

Poehner, Matthew E., and James P. Lantolf. 2005. "Dynamic Assessment in the Language Classroom." *Language Teaching Research* 9(3):233-265.

Read, John, and Belinda Hayes. 2003. "The Impact of IELTS on Preparation for Academic Study in New Zealand." *IELTS Research Reports* 4:153-205.

Saville, Nick. 2012. "Applying a Model for Investigating the Impact of Language Assessment within an Educational Context: The Cambridge ESOL Approach." British Council/IDP Australia IELTS.

Scarino, Angela. 2013. "Language Assessment Literacy as Self-awareness: Understanding the Role of Interpretation in Assessment and Teacher Learning." *Language Testing* 30(3):309-327. Research Reports volume 50:4-8. Accessed March 10, 2019. http://www.ielts.org/researchers/research.aspx.

Shih, Chih-Min. 2010. "The Washback of the General English Proficiency Test on University Policies: A Taiwan Case Study." *Language Assessment Quarterly* 7:234-254.

Shohamy, Elana. 2001. "Democratic Assessment as an Alternative." *Language Testing* 18:373-391.

Spada, Nina, and Maria Fröhlich. 1995. *Communicative Orientation of Language Teaching Observation Scheme. Coding Conventions and Applications.* Sydney: NCELTR, Macquarie University Stake.

Stiggins, Richard J. 2001. "The Unfulfilled Promise of Classroom Assessment. Educational Measurement." *Issues and Practice* 20(3):5-15.

Tsagari, Dina. 2009. "Revisiting the Concept of Test Washback Investigating FCE in Greek Language Schools." *Research Notes* 35:5-10. Accessed March 10, 2019. http://www.cambridgeenglish.org/images/23154-research-notes-35.pdf.

—. 2011. "Washback of a High-stakes English Exam on Teachers' Perceptions and Practices." Accessed March 10, 2019. http://www.enl.auth.gr/symposium19/19thpapers/042_Tsagari.pdf.

—. 2012. "FCE Exam Preparation Discourses: Insights from an Ethnographic Study." *Research Notes* 47:36-48. Accessed March 10, 2019. http://www.ielts.org/researchers/research.aspx.

Tsagari, Dina, and Vogt, Karin. 2017. "Assessment Literacy of Foreign Language Teachers around Europe: Research, Challenges and Future Prospects". *Papers in Language Testing and Assessment* 6(1):41-63.

Vogt, Karin, and Dina Tsagari. 2014. "Assessment Literacy of Foreign Language Teachers: Findings of a European Study." *Language Assessment Quarterly* 11:374-402.

Vogt, Karin, Ines Sperling, and Moritz Brüstle. 2018. "TALE–Teachers' Assessment Literacy Enhancement Needs Analysis Report." Erasmus+ European Union. Project No. 2015-1CY01-KA201-011863 (Project No. 2015-1-CY01-KA201-011863). Accessed July 20, 2019. http://taleproject.eu/mod/page/view.php?id=1198.

Wall, Dianne, and Charles J. Alderson. 1993. "Examining Washback; The Sri Lankan Impact Study." *Language Testing* 10(1):41-69.

Wall, Dianne. 1997. "Impact and Washback in Language Testing." In *Encyclopedia of Language and Education, Vol 7. Language Testing and Assessment,* edited by Caroline Clapham, and David Corson, 291-302. Dordrecht: Kluwer Academic.

—. 2000. "The Impact of High-stakes Testing on Teaching and Learning: Can This Be Predicted or Controlled?" *System* 28:499-509.

Wallace, Michael J. 2000. *Action Research for Language Teachers*. Cambridge: Cambridge University Press.

Watanabe, Yoshinori. 1997. "The Washback Effects of the Japanese University Entrance Examinations of English-classroom-based Research." PhD dissertation. University of Lancaster.

—. 1996. "Does Grammar-Translation Come from the Entrance Examination? Preliminary Findings from Classroom-based Research." *Language Testing* 13(3):318-333.

—. 2000. "Washback Effects of the English Section of Japanese University Entrance Examinations on Instruction in Pre-college Level EFL." *Language Testing Update* 27:42-47.

—. 2008. "Methodology in Washback Studies." In *Washback in Language Testing; Research Context and Methods*, edited by Liying Cheng, Yoshinori Watanabe, and Andy Curtis, 19-36. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Yin, Muchun. 2010. "Understanding Classroom Language Assessment through Teacher Thinking Research." *Language Assessment Quarterly* 7:175-194.

Yin, Robert K. 2003. *Case Study Research: Design and Methods*. Thousand Oaks: Sage.

Zhang, Cong, and Yan, Xun. 2018. "Assessment Literacy of Secondary EFL Teachers: Evidence from a Regional EFL Test." *Chinese Journal of Applied Linguistics* 41(1):25-46.

CHAPTER FOURTEEN

THE CHALLENGE OF ALIGNING VOCABULARY AND GRAMMAR DIFFICULTY LEVELS WITH CEFR PROFICIENCY RATINGS: THE CASE OF THE E-CAT

JACK BURSTON, ANDROULLA ATHANASIOU AND MARO YIOKARI NEOPHYTOU

This study focuses on how the *English Profile* can further the language assessment literacy (LAL) of L2 English teachers by providing an objective means of aligning test questions with *Common European Framework of Reference* (CEFR) levels. It does this with regard to a particular test, the English E-CAT, but the outcomes of the study are applicable to any test.

By design, because it is intended to be applicable to any language, the CEFR indicates very little about the specific vocabulary and grammar that occur at any competence level. The CEFR has left it to external stakeholders to correlate specific linguistic features with proficiency levels for particular languages. Recently, the *English Profile* has done precisely this with the public release of its *English Vocabulary Profile* (EVP) and *English Grammar Profile* (EGP).

The E-CAT, a computer-adaptive test of English, streams students into four levels based on responses to questions of known difficulty levels. In order to align the E-CAT with the CEFR, the four levels of the E-CAT were equated with A2, B1, B2 and C1. This was done subjectively on the basis of previous teaching experience.

The *English Profile* has now made it possible to objectively verify the assignment of CEFR rankings of the E-CAT difficulty levels. Specifically, within this study three research questions are addressed:

1. To what extent do the difficulty levels of the E-CAT questions objectively correspond to the CEFR ratings that have been assigned to them?
2. To what extent can these CEFR grammatical ratings provide an objective basis for fixing the cut-off points between statistically determined E-CAT difficulty levels?

3. What grammatical features are criterial in defining CEFR levels?

The answers to these questions have not only shed useful light on the E-CAT, but also considerably enhanced the LAL regarding the *English Profile* and its applicability as a tool in determining the CEFR alignment of test questions in general.

1. Introduction

One of the most critical aspects of language assessment literacy (LAL) is being able to objectively match assessment to the competence level of learners. This is particularly so when designing a proficiency test that is independent of any specific curriculum. Computer-adaptive testing offers a viable solution to this challenge in that assessment is based on an inventory of test items that cover a wide range of difficulty levels. The creation of a computer-adaptive test requires two components, a question inventory of known difficulty levels and a question presentation algorithm. Fortunately, a freely available, easy-to-use, Computer-Adaptive-Test template, the *St Louis University Placement Exam* (*SLUPE*), existed to handle question presentation. All instructors needed to do was supply the multiple-choice test questions, tagged with a difficulty level of 1 (the easiest) to 4 (the most difficult).

2. The E-CAT

At the Cyprus University of Technology, members of the Language Centre used *SLUPE* to create an L2 English proficiency test known as the E-CAT (Burston and Neophytou 2014; Burston, Neophytou and Lamprianou 2016) for incoming first-year students. *SLUPE* generates computer-adaptive English language tests based on item response theory, which presupposes that proficiency in a given domain can be measured by the difficulty level of questions that can be correctly answered. The difficulty level of questions is determined through statistical analysis of actual student responses, which in the case of the E-CAT was established with considerable precision on the basis of nearly 1200 administrations of the test. The E-CAT automatically adjusts to the proficiency level of students by presenting easier questions following incorrect responses and more difficult ones after correct answers. Ultimately, the algorithm zeroes in on the difficulty level at which correct responses are most consistent.

The E-CAT consists of 193 multiple-choice questions of which 102 are text-based, 65 audio-based and 26 video-based. These cover vocabulary, grammar and contextually appropriate responses to prompts. The questions comprise 1483 complete sentences totalling nearly 13,000 words. Based on a difficulty scale that extends from -3.50 (the easiest) to +3.50 (the most difficult), test questions were assigned to one of four difficulty levels. Questions with a difficulty rating less than -1 were classified as E-CAT level 1, those from -1 to less than 0 as E-CAT level 2, those from 0 to less than +1 as E-CAT level 3 and at +1 and above as E-CAT level 4. Furthermore, the difficulty levels 1-4 of the test were aligned with *Common European Framework of Reference* (CEFR) proficiency rankings. However, unlike the statistical determination of question difficulty levels, this was done without any objective way of validating the correspondence.

3. The CEFR

The *Common European Framework of Reference* (CEFR) was developed to provide an explicit and generic means of describing foreign language competency, which it does from least to most advanced in terms of three major levels (A, B, C), each of which consists of two sublevels (A1/A2, B1/B2, C1/C2). In order to align the E-CAT with the CEFR, its four difficulty levels were equated with A2-C1 on the basis of previous teaching experience. In the CEFR, linguistic competence is defined relative to what learners actually “can do” with the language. That is to say that it is performance-based. For example, at B1 level the learner: *Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans* (Council of Europe 2001, 24).

By design, because it is intended to be applicable to any language, the CEFR indicates very little about the specific vocabulary and grammar that occur at any competence level. Descriptors include only very general reference to the vocabulary expected of learners performing certain skills, for example, involving “familiar words” and “high frequency, everyday job-related language”. The same holds true with regard to syntactic parameters, which are described in terms such as “very basic phrases”, “simple connected text”, etc. The CEFR has left it to external stakeholders to correlate specific linguistic features with proficiency levels for particular languages.

Early studies relating to the CEFR focused attention on the vocabulary associated with performance at some of the CEFR levels (Coste et al. 1987). At issue was vocabulary felt to be necessary at certain CEFR levels based on general word frequency lists. The grammatical correlates of

CEFR levels for English were first investigated by Van Ek and Trim (1991). This issue was later treated at considerable length by Hawkins and Filipovic (2011), who based their work on an analysis of an early version of the Cambridge Learner Corpus. More recently, the *English Profile* has taken this effort much further with the public release of its *English Vocabulary Profile* (EVP) and *English Grammar Profile* (EGP).

4. The English Profile

The *English Profile* is a continuation and significant expansion of previous efforts to correlate vocabulary and grammar with CEFR proficiency levels in English. It consists of two online searchable databases, the *English Vocabulary Profile* (EVP) and the *English Grammar Profile* (EGP). The *English Profile* was developed by a European Lifelong Learning consortium, the principal members of which are the Cambridge University Press and the Cambridge English for Speakers of Other Languages (ESOL) testing service. The EGP derives primarily from the Cambridge Learner Corpus, which is extracted from the written scripts of over 333,000 exam responses from the various Cambridge ESOL tests at all CEFR levels. The EVP also draws upon the Cambridge English Corpus, a billion-word compilation of written and spoken English language from multiple sources.

Reversing the normal process of using standardized tests to determine the proficiency level of language learners, the *English Profile* uses the demonstrated proficiency level of ESOL test takers to determine the corresponding CEFR level of the vocabulary and grammatical constructions within test responses. This however, comes with two caveats. Firstly, one has to assume the accuracy of the underlying student proficiency ratings. However, this is not an unreasonable assumption given the well-earned professional testing status of the Cambridge exams and the extensiveness of the *English Profile* databases. Secondly, the proficiency level at which L2 English vocabulary and grammatical constructions occur in student productions can be very much influenced by the L1 of the test taker. For example, because of the considerable number of cognates in English and French, an English A2 level test taker whose native language is French is much more likely to use more advanced level English abstract nouns like "impression (B2/C2), sincerity (C2), justice (B2/C2)" than an L1 Chinese or Arabic A2 level English learner for whom such cognates do not exist. Likewise, the proper use of articles and determiners would come much more readily to A2 level learners whose L1 uses these grammatical forms than to those whose L1 does not.

Notwithstanding, the extensive size of the *English Profile* databases and the very broad range of L1 test takers mitigate against this bias. In sum, then, it is not unreasonable to accept at face value the CEFR proficiency level rankings of the vocabulary and grammar of the EVP and EGP databases.

5. English Vocabulary Profile

The EVP, which gives selectable access to both British and American varieties, contains some 20,800 headwords. Searches can be made either globally, i.e., including all CEFR levels A1-C2, or by specific levels or level ranges (e.g., B1 only, A2-B1). It can be browsed alphabetically or by typing in a specific word (including a wildcard character). A particularly notable feature of the EVP look-up is its context sensitivity. Depending on its contextual meaning, the same headword may be classified in multiple ways anywhere from A1 to C2 (see Figure 1 for "feel").

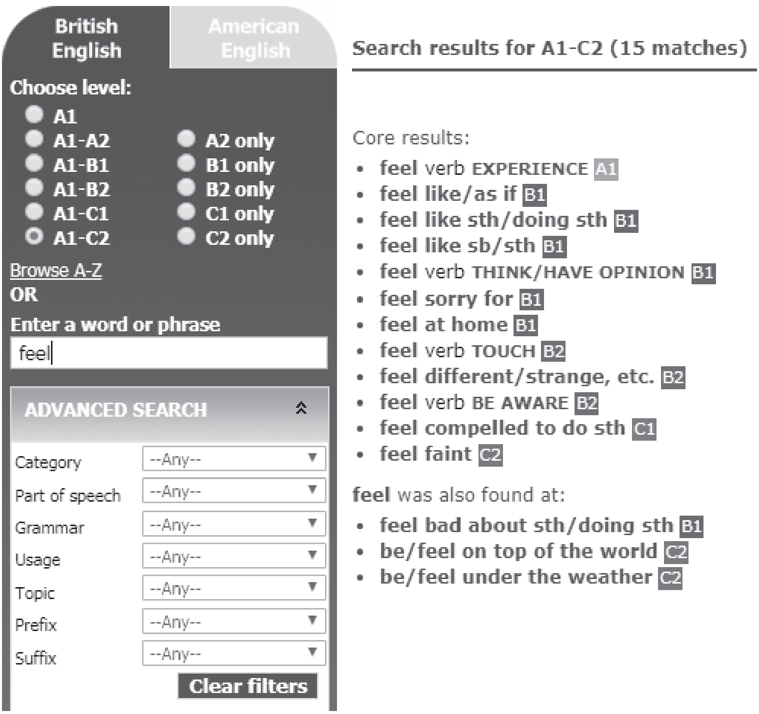


Figure 1: English Vocabulary Profile

A number of advanced search options allow the EVP look-up to be further refined by category (words, phrases, phrasal verbs, idioms), part of speech, grammar (plurals, count/non-count, transitive/intransitive), usage, topic, prefix, and suffix. These options are especially useful when evaluating a pedagogical resource for its suitability for a specific learner proficiency level.

6. English Grammar Profile

The EGP database consists of 1222 grammatical descriptors organized in alphabetical order and searchable by any combination of CEFR levels (Figure 2). Each descriptor is accompanied by student example sentences and test details (e.g., the year the test was taken, the L1 of a student, etc.).

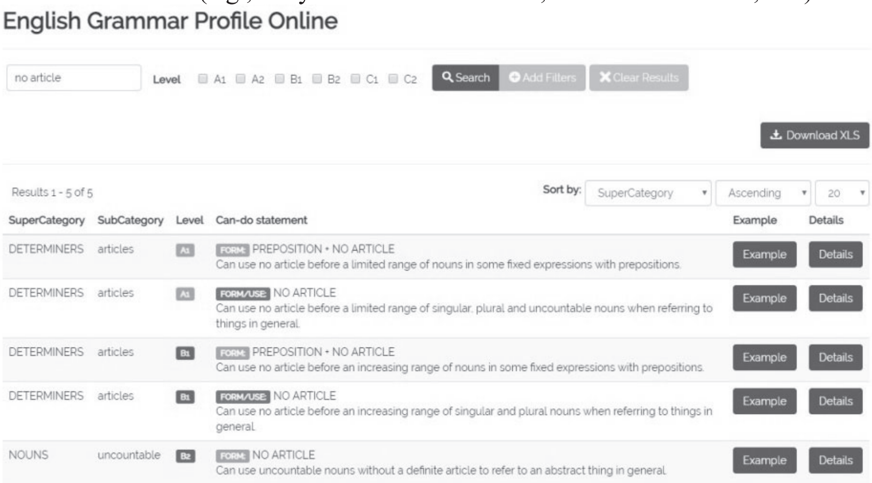


Figure 2: English Grammar Profile

The number of grammatical descriptors at each level varies from 109 to 338, with the greatest number in the A2-B2 range (Table 1).

A1	109
A2	291
B1	338
B2	243
C1	129
C2	112

Table 1: CEFR Grammatical Descriptors

7. Research questions

This study of the *English Profile* focuses on determining its usability in correlating the difficulty levels of E-CAT test items with CEFR proficiency ratings. Using the EVP and EGP, the following three research questions are posed:

1. To what extent do the difficulty levels of the E-CAT questions objectively correspond to the CEFR ratings that have been assigned to them?
2. To what extent can these CEFR grammatical ratings provide an objective basis for fixing the cut-off points between statistically determined E-CAT difficulty levels?
3. What grammatical features are criterial in defining CEFR levels?

8. E-CAT/CEFR Vocabulary Analysis

Despite the relatively small size of its database, excluding proper names and numbers, the EVP was able to locate about 96% of the vocabulary at all E-CAT levels. In using the EVP, it needs to be recognized that, in all but the simplest texts specifically contrived to be exclusively at the A1 level, a range of vocabulary levels is bound to occur. What characterizes the lexical level of a text is the relative proportion of vocabulary at each CEFR level, with the basic A1/A2 level vocabulary predominant at all levels. This is clearly attested in the EVP results from the lexical analysis of the E-CAT (Table 2). As can be seen, even at E-CAT level 4, well over 70% of the vocabulary remains at the A1-A2 level.

CEFR Vocabulary Levels

CEFR	E-CAT Level 1 words		E-CAT Level 2 words		E-CAT Level 3 words		E-CAT Level 4 words	
A1	1129	72.23%	3261	61.75%	1518	52.78%	1501	52.61%
A2	279	17.85%	798	15.11%	601	20.90%	577	20.22%
B1	108	6.91%	642	12.16%	410	14.26%	445	15.60%
B2	33	2.11%	409	7.74%	274	9.53%	234	8.20%
C1	8	0.51%	120	2.27%	73	2.54%	96	3.36%
C2	6	0.38%	51	0.97%	52	2.54%	29	3.36%
	1563		5281		2876		2853	
A1/2		**90.08%**		**76.86%**		**73.68%**		**72.84%**

Table 2: CEFR Vocabulary Levels at E-CAT Levels 1-4

Notwithstanding, there is a very substantial increase in the proportion of B1-B2 level vocabulary between E-CAT levels 1 and 2, which more than doubles (from 9.02% to 19.9%), and again between E-CAT levels 2 and 3 which increases by a further 19% (from 19.9% to 23.79%) while remaining stable at E-CAT level 4. Similarly, the proportion of C1-C2 vocabulary more than triples between E-CAT levels 1 and 2 (from .89% to 3.24%), and again between E-CAT levels 2 and 3 which increases by nearly a further 57% (from 3.24% to 5.08%) as well as another 32% (from 5.08% to 6.72%) between E-CAT levels 3 and 4. Thus, as regards vocabulary, the EVP analysis clearly confirms the correlation of E-CAT levels 1-4 with CERF ratings of A2-C1.

9. E-CAT/CEFR Grammar Analysis

Although the vocabulary in E-CAT levels 1-4 demonstrably corresponds to their assigned CEFR A2-C1 rankings, as indicated in the *English Profile* booklet (12), CEFR competence levels are held to be primarily determined by syntactic rather than lexical features. Accordingly, the EGP must be considered the final arbitrator of the correlation between E-CAT difficulty levels and CEFR ratings. That being said, however, it needs to be kept in mind that the CEFR ranking of grammatical constructions is very contextually sensitive to their lexical constituents. As can be seen in Figure 2, the same construction (e.g., no article) can be classified as anything from A1 to B2 depending on whether it occurs with a "limited, extended, or extensive" range of vocabulary items. Sometimes specific vocabulary is listed, but most often only illustrative examples are indicated. In this case, the EVP needs to be consulted to determine the CEFR level of constituent vocabulary to establish the CEFR ranking of a particular instance of a grammatical construction.

While determining the CEFR level of the vocabulary of sentences within the text/audio/video-based questions of the E-CAT is very straightforward using the EVP, performing the same operation within the EGP for grammatical constructions is exceedingly laborious and time-consuming owing to the very large number of grammatical descriptors, 1222 in total. The task is made all the more difficult because the grammatical classifications of the EGP do not always conform to traditional terminology. For example, searching for "present participle" or "gerund" will not find anything. Such forms can only be located by looking for "-ING", which includes all verbal forms with this ending (e.g., tenses and verbal complements). As a consequence, searching the EGP often requires reading through dozens of grammatical descriptors that at first glance do not appear related to the constructions that are sought. Aware of the difficulty posed by having to work with such a large inventory of grammatical identifiers, the *English Profile* booklet (16-24) proposes a much more manageable set of 48 criterial, i.e., defining, grammatical features for CEFR levels A2-C2 (Table 3).

CEFR Level	Number of Criterial Features	Total Number of Descriptors
A2	12	291
B1	18	338
B2	10	243
C1	5	129
C2	3	112
	48	1103

Table 3: *English Profile* Criterial Grammatical Features

Unfortunately, however, these criterial features raise a serious question of justification and even more so of application. Firstly, although the sources of these features are acknowledged, most notably Hawkins and Filipovic (2011), it is never explained on what basis this small number of grammatical categories, of the dozens at each CEFR level in the EGP, were selected. In fact, the selection of criterial features is largely based on the work of Van Ek and Trim (1991, 2001). However, the closest these authors come to any objective measurements is what they call their 10-to-1 rule, which determines the level at which a (previously assumed criterial) grammatical feature is placed. If feature X occurs 12 times at level A2 and 100 times at level B1, it is deemed to be criterial for level A2. To be considered criterial at level B1, it would need to occur at that level at least 120 times. Secondly, even if there is good reason for selecting just the features proposed by the *English Profile*, it is difficult to apply them when searching the EGP. For example, traditional grammatical terminology used in the criterial descriptions such as “intransitive”, “infinitival complement”, “post-nominal modification”, “genitive” as well as more transformational linguistic oriented descriptors like “Tough-Movement”, “Subject-to-Subject-Raising”, “Subject-to-Object-Raising”, and “Pseudo-cleft” are not to be found in the EGP. It is only possible to locate within the EGP exact matches for 16 of the 48 purported criterial features (7/12 A2, 6/18 B1, 1/10 B2, 2/5 C1, 0/3 C2). Similar, but not identical, matches can be found for an additional 16 (1 A2, 6 B1, 5 B2, 2 C1 and 2 C2). The remaining 16 simply do not exist within the EGP. These discrepancies stem from the fact that the identification of CEFR criterial grammatical features was undertaken before the publication of the *English Profile* and so does not take account of the results of the EGP.

While the principle of correlating CEFR levels with criterial grammatical features is well founded, this manifestly cannot be done on the basis of the features indicated in the *English Profile* booklet. In the absence of any principled way to distinguish between criterial and non-criterial grammatical features, in analyzing the E-CAT questions there was no alternative but to take note of all categories that could be found through a comprehensive search of the EGP. The overall results of this initial analysis were quite disparate, with numerous repetitive and overlapping categories. For example, using all of the EGP descriptors, the grammatical analysis of the sentence "Pollution from coal-powered factories turned the cities black" is shown in Table 4.

Grammatical Feature	CEFR Level
Affirmative declarative	A1
Past simple ordering of events	B1
Noun subject	A1
Determiner no article + noun	B1
Uncountable noun	B1
Noun + prepositional phrase	B1
Preposition	A2
Compound-adjective	C1
Adjective + plural noun	A1
Attributive adjective + noun	B1
Determiner no article + noun	B1
Regular plural -IES	A2
Linking verb + complement	B2
Determiner + noun	A1
Regular plural -IES	A2
Predicative adjective	A2

Table 4: Unintegrated EGP Grammatical Feature Analysis

By grouping the sentence constituents into more inclusive grammatical structures, it was possible to produce the much more cohesive analysis shown in Table 5.

Grammatical Features	**Category**	**CEFR Level**
Affirmative declarative past simple ordering events	Sentential	B1
Determiner no article + uncountable noun subject	Noun phrase	B1
Noun + preposition + determiner no article + compound adjective + plural noun -IES	Complex noun phrase	C1
Linking verb + determiner + plural noun -IES + adjective complement	Verb phrase	B2

Table 5: Integrated EGP Grammatical Feature Analysis

This reanalysis firstly involved reducing the many redundancies in the EGP to single categories. For example, the twelve different uses of the word WOULD at the B1 level were simply treated as B1 modal WOULD. Since many of the EGP descriptors are quite long (e.g., IT' + 'BE' + ADJECTIVE + 'FOR' + OBJECT + INFINITIVE), where necessary these were shortened to facilitate their tabulation in a spreadsheet. Syntactic simplification involved incorporating sentential features into a single category that included grammatical categories relating to the affirmative/ negative, declarative/interrogative, and active/passive nature of the utterance. Also included in the sentential category was the occurrence of modals and the nature of the verbal tense. Similarly, the constituents of major phrase types (i.e., noun phrases, verb phrases, prepositional phrases) were combined into single comprehensive categories. So, too, where relevant, grammatical functions such as subject, object, and complement were included in these categories. As can be seen, since lower level CEFR features persist at higher levels, these major categories all contain elements at various CEFR levels. For example, the constituents of the complex noun phrase “pollution from coal-powered factories” (noun + preposition + determiner no article + compound adjective + plural noun -IES) are shown in Table 6.

Grammatical Feature	CEFR Level
Noun + prepositional phrase	B1
Preposition	A2
Determiner no article + noun	B1
Compound adjective	C1
Plural noun -IES	A2

Table 6: Criterial Feature Determination Example

By definition, lower level features cannot be criterial for higher level constructions, so it is the highest level feature which determines the ranking of the entire grammatical construction, C1 in this case of the example sentence. On the basis of this identification of comprehensive grammatical categories and criterial features within them, it was possible to not only simplify and unify the analysis of the E-CAT but also ultimately to define a set of criterial features that may usefully serve for future analyses of other tests and, more generally, other pedagogical materials. This will be discussed in the last part of this chapter.

In analyzing the grammatical constituents of any linguistic document, whether it be a test like the E-CAT, a student composition, or a teaching resource such as a reading passage or an audio/video clip, the broad range of CEFR levels within it becomes immediately apparent. As with vocabulary, even in the simplest text, unless it has been specifically prepared to only contain A1 level constructions, more advanced level grammatical categories are bound to appear. Likewise, basic structures classified as A1 such as noun or pronoun subjects or objects, determiner + noun, regular plural nouns, simple present tense, simple past tense, the verb BE and a complement, etc., occur in texts at all levels. So, too, texts above the A1 level will inevitably contain lower and higher level constructions. The *English Profile* explicitly recognizes this heterogeneity when it states that the grammatical features of one CEFR level are presumed to persist at higher ability levels (9). So, too, it is not the occurrence of any one particular construction that determines the CEFR level of a pedagogical resource, but rather the relative proportion of features at given CEFR levels. This is clearly shown in the results of the analysis of defining CEFR grammatical features within the four levels of the E-CAT (Table 7).

E-CAT Level 1 Grammar Constructions			E-CAT Level 2 Grammar Constructions		E-CAT Level 3 Grammar Constructions		E-CAT Level 4 Grammar Constructions	
A1	584	53.82%	1547	43.99%	667	37.68%	692	36.44%
A2	325	29.95%	813	23.12%	481	27.18%	541	28.49%
B1	157	14.47%	847	24.08%	482	27.23%	465	24.49%
B2	16	1.47%	246	6.99%	118	6.67%	145	7.64%
C1	3	0.28%	48	1.36%	17	0.96%	56	2.95%
C2	0	0.00%	16	0.45%	5	0.96%	2	2.95%
	1085		3517		1770		1899	
A1/A2		**83.78%**		**67.10%**		**64.86%**		**64.93%**

Table 7: CEFR Grammar Rankings at E-CAT Levels 1-4

As was the case with vocabulary, A1-A2 grammatical features represent the majority of categories at all E-CAT levels and even at level 4 they constitute nearly 65% of all constructions. The proportion of B1-B2 categories nearly doubles between E-CAT levels 1 and 2 (from 15.94% to 31.07%), with another 9% (31.07% to 33.9%) increase between E-CAT levels 2 and 3. The percentage of B1-B2 categories remains virtually unchanged between E-CAT levels 3 and 4, which is due to the fact that there is a three-fold increase (from 1.92% to 5.9%) between levels 3 and 4 at C1-C2. Between E-CAT levels 1 and 2, the proportion of C1-C2 grammatical constructions increases six-fold (from .28% to 1.81%) and another 6% (from 1.81% to 1.92%) between E-CAT levels 2 and 3. While the absolute number of C1-C2 constructions is small at all E-CAT levels, their presence along with the sizeable increases of B1-B2 constructions, very much confirms the correlation between E-CAT levels 1-4 and the CEFR A2-C1 ratings assigned to them.

10. E-CAT Level Cut-off Points

While, overall, the correlation between E-CAT 1-4 difficulty levels conforms to their assigned CEFR A2-C1 rankings, the question remains as to the accuracy of the cut-off points between the levels. As previously mentioned, these were initially set without the benefit of being able to objectively confirm the CEFR rankings of the test questions. In order to determine the correlation between the E-CAT difficulty level cut-off points and CEFR rankings for grammatical constructions, five questions were checked in the immediate vicinity (above and below) of the E-CAT level 2/3/4 cut-off points (-1.00, 0.00, +1.00). Where the proportion of B1+ grammatical constructions was substantially above or below the average for the level, these questions were accordingly raised or lowered a level. This readjustment was subject to being within the standard error of the questions' difficulty level. For example, an E-CAT level 1 question with a difficulty of -1.047 (i.e., close to the E-CAT level 2 cut-off of -1.00) with a standard error of 0.14 could have a true value as high as -0.907, so moving it up a level remains within the statistical accuracy of its difficulty level.

At E-CAT levels 1 and 2, the results very much confirmed the accuracy of the -1.00 and 0.00 cut-off points, with only a few questions needing to be reassigned. Specifically, among the most difficult E-CAT level 1 questions (i.e., those closest to -1.00), four contained grammar features above the B1+ average (from 33.34% to 50% compared to 31.07%) for E-CAT level 2 and were thus reassigned to E-CAT level 2. When the easiest questions at E-CAT level 2 (again those closest to -1.00), were examined, three were found to have B1+ grammar constructions well below the average (from 0% to 7.69% compared to 31.07%) for E-CAT level 2 and were lowered to E-CAT level 1. At the upper end of E-CAT level 2 (i.e., closest to 0.00), four questions contained B1+ grammatical features that ranged from 31.93% to 75%, compared to the E-CAT level 3 average of 35.14% and were thus reassigned to E-CAT level 3. Among the easiest questions at E-CAT level 3 (i.e., closest to 0.00), three had B1+ grammatical features that ranged from 16.67% to 28.57%, below the E-CAT level 3 average of 35.14%. Accordingly, these were moved down to E-CAT level 2. At the top end of E-CAT level 3, one question contained 75% B1+ grammatical constructions, more than double the 35.14% average for that level and was therefore raised to E-CAT level 4. There were, however, a number of other discrepancies within E-CAT levels which are discussed below. Lastly, three of the easiest questions (i.e., closest to 1.00) at E-CAT level 4 contained B1+ grammatical constructions considerably below the average (from 20% to 25% compared to 35.07%) for E-CAT level 4 and were reassigned to E-CAT level 3.

11. E-CAT Level 3 Discrepancies

In comparing the distribution of CEFR features within E-CAT level 3, it was discovered that, despite having difficulty levels above the 0.00 cut-off point of between 0.225 and 0.82, five questions contained no grammatical constructions above A2. Two of these questions tested lexical expressions. One (beat around the bush) was identified in the EVP as C1. The other (the creeps) was not found in the EVP and, based on the fact that it does not occur within the top 60,000 words of the Corpus of Contemporary American English, would need to be considered at least at C1 level. The third question tested the use of the subject pronoun "all" requiring a verb in the singular, an A1 level feature in English. The corresponding pronoun in Greek always accords in the plural, making it a very frequent source of error for native Greek speakers even for advanced L2 English learners. The last two questions required test takers to select from one to five appropriate responses to a prompt, with all of the answers involving only A1/A2 grammatical constructions. In sum, then, the discrepancies between the relatively low CEFR grammatical features and the high level of difficulty of these ECAT level 3 questions demonstrate that factors other than the L2 English grammar level can, and in fact do, determine the difficulty level of questions. Specifically, if targeted vocabulary is not known by a large number of test takers then, however simple the grammar may be, a test item will statistically prove to be difficult. Likewise, if a low-level grammatical construction in English runs counter to what is required in the L1 of a test taker, it is likely to be frequently missed and thus have a relatively high difficulty rating. Lastly, context plays a critical role in the understanding of any text, especially in regard to connotations of which less advanced-level language learners may not be aware.

12. Criterial Grammatical Features

Using the inclusive grammatical categories previously described, a total of 422 distinct types covering a total of 11,950 tokens emerged from the analysis of E-CAT levels 1-4. It being a reasonable assumption that the most frequently occurring grammatical categories are those which have been the most firmly acquired, and thus the most criterial, a statistical analysis of the relative frequency of occurrence of the CEFR A1-C2 grammatical descriptors within the E-CAT was undertaken. At all E-CAT levels, there was a broad distribution of CEFR rankings, with the frequency of occurrence of the 422 distinct types ranging from 0.10% to

24%. The top 2-4 most frequent construction types within each CEFR ranking typically occurred about 15% of the time and another 2-4 about 5%-9% of the time. On average, there were about 11 grammatical construction types with a frequency of occurrence above 2% and less than 5%. The great majority (334/422, 79%) of grammatical construction types at all CEFR rankings occurred less than 2% of the time. Over 60% of these (205/334) had a frequency of occurrence below 1%. As can be seen, there is an inverse relationship between the frequency of occurrence of grammatical construction types and the number of grammatical construction types at any cut-off point. The lower the frequency of occurrence, the greater is the number of distinct construction types.

The guiding principle in determining a cut-off point for identifying the most frequent, and thus criterial, grammatical features was to end up with an inventory representative of the most frequent constructions yet of manageable size for future application purposes. Four cut-off points among the most frequent construction types were tried: 2%+, 3%+, 4%+, and 5%+. The first resulted in 88 grammatical construction types, the second in 65, the third in 40, and the fourth in 32. The results are shown in Table 8.

Frequency of Occurrence	Construction Types	Total Tokens
5%+	32	5552
4%+	40	6490
3%+	65	7705
2%+	88	8900

Table 8: Frequency of Occurrence of E-CAT CEFR Descriptors

The 3%+ distribution resulted in the best combination of maximum coverage (7705 tokens out of 11,950, 65%) with a manageable set of criterial features. Moreover, the 65 criterial features for CEFR level A1-C2 constructions also most closely matched the number proposed by the *English Profile*, which identified 48 for CEFR levels A2-C2, i.e., omitting A1. The internal distribution of E-CAT grammatical criterial features resulting from the 3%+ cut-off point is shown in Table 9.

CEFR Level	Total Grammatical Constructions	Most frequently Occurring Grammatical Constructions (3+%)
A1	50	11
A2	109	11
B1	139	8
B2	76	9
C1	34	12
C2	14	14
Total	**422**	**65**

Table 9: E-CAT Criterial Grammatical Constructions

The total number of criterial features is thus 65 out of a total 422, i.e., about 15%. This compares to the *English Profile* estimate of 48 out of 1103, i.e., about 4%, but which does not include A1 level constructions. Though higher than that proposed by the *English Profile,* the E-CAT calculations are based on the analysis of actual frequency of occurrence and are thus objectively verifiable whereas those proposed by the *English Profile* are not. More importantly, unlike the latter, the grammatical descriptors of the E-CAT derive directly from the categories used in the EGP. The actual grammatical categories included in the criterial constructions of the E-CAT are shown in Table 10.

E-CAT Criterial Grammatical Constructions	Frequency of Occurrence		E-CAT Criterial Grammatical Constructions	Frequency of Occurrence
A1 declarative	16.08%		A2 present simple verb	6.92%
A1 pronoun subject	14.78%		A2 irregular plural noun	6.82%
A1 present simple	11.65%		A2 subordinating conjunction	5.86%
A1 determiner + noun	9.78%		A2 uncountable noun	5.76%
A1 noun subject	5.83%		A2 preposition	5.53%
A1 preposition + noun phrase	4.77%		A2 predicative adjective	5.49%
A1 coordinating conjunction	4.68%		A2 adverb at front	3.92%
A1 noun object	4.61%		A2 adverb time	3.55%
A1 proper noun	3.95%		A2 mental process verb	3.37%
A1 regular plural noun	3.89%		A2 generic YOU	3.18%
A1 past simple	3.34%		A2 WH- affirmative interrogative	3.00%
B1 complex noun phrase	16.23%		B2 no article + noun	24.02%
B1 attributive adjective	16.18%		B2 generic WE/US	7.87%
B1 no article + noun	15.42%		B2 adverb manner	6.00%
B1 past simple verb	6.93%		B2 preposition	6.00%
B1 past simple order	5.61%		B2 adjective + to + infintive	4.76%
B1 uncountable noun	4.60%		B2 verb + particle + object	4.55%
B1 present simple verb	3.69%		B2 adverb degree	3.93%
B1 past simple habitual	3.08%		B2 adverb time	3.73%
			B2 phrasal verb no object	3.73%
C1 noun + past participle	13.33%		C2 IT cleft THAT	14.81%
C1 no article + noun	11.85%		C2 NEITHER verb inversion	14.81%
C1 uncountable noun	8.15%		C2 THIS immediacy	14.81%
C1 WHAT cleft + BE	7.41%		C2 verb + obj + part + prep	11.11%
C1 subordinating conjunction	5.93%		C2 subordinating conjunction	7.41%
C1 -ING abstract noun	5.19%		C2 WH- cleft focus	7.41%
C1 preposition	4.44%		C2 adverb stance	3.70%
C1 reporting verb	4.44%		C2 adverb time	3.70%
C1 adverb + comparative	3.70%		C2 BE NOT THAT MUCH	3.70%
C1 adverb degree	3.70%		C2 HAVE + object + inf passive	3.70%
C1 GET/HAVE + object + -ING	3.70%		C2 HOW + adverb	3.70%
C1 OTHERS	3.70%		C2 modal HAVE TO	3.70%
			C2 noun's possessive ellipsis	3.70%
			C2 THESE in narratives	3.70%

Table 10: Criterial Grammatical Features of the E-CAT

13. Conclusions and Study Limitations

To conclude, by incorporating the hundreds of individual CEFR descriptors into more inclusive grammatical constructions, it is possible to greatly reduce the number of categories that have to be evaluated and to precisely identify the determining criterial features within them. In so doing, it has been possible to apply the *English Profile* CEFR ratings for vocabulary and grammar features to empirically confirm the overall correlation between the four difficulty levels of the E-CAT and the A2/C1 ratings that had been assigned to them. So, too, an analysis of the relative percentages of CEFR rankings in the grammar of questions on the borderline between E-CAT levels has allowed their more accurate reassignment to an appropriate level. It also demonstrated that a question's difficulty level may be determined by factors other than grammar, e.g., vocabulary, L1 interference, and contextual connotations. Lastly, the relative frequency of occurrence of specific grammatical constructions at CEFR levels A1-C2 has resulted in the identification of a small, yet representative, set of 65 criterial features that have the potential to be used in determining the CEFR grammatical features of other tests. From the perspective of furthering language assessment literacy, this is arguably the most important outcome of this study, for without a manageable set of criterial grammatical features, searching the EGP is a very slow and tedious undertaking. This very much detracts from the great potential usefulness of the EGP as a tool to be used in conjunction with the EVP for correlating CEFR levels with the vocabulary and grammatical constructions of tests and other pedagogical resources.

Despite the successful results of the application of the *English Profile* to the analysis of the E-CAT, there are some important limitations to this study. Firstly, it needs to be kept in mind that the CEFR rankings of the vocabulary and grammar within the *English Profile* databases derive from active language production. There is thus an important difference when they are used to correlate CEFR levels within texts, like the reading, audio, or video passages of the E-CAT, which primarily involve receptive linguistic skills. It being more difficult to produce than to recognize language at a given level, CEFR correlations in such cases are likely to overestimate the actual difficulty level of a resource. Secondly, while providing a step in the right direction by demonstrating that grammatical criterial features can be objectively identified through a statistical analysis of their frequency of occurrence, this study is limited by the relatively small database of the E-CAT, which only consists of some 1483 sentences containing about 13,000 words. To maximize the usefulness of the EGP, what is really needed is for the *English Profile*, with its massive *Cambridge Learner Corpus* and *Cambridge English Corpus* databases, to be similarly analyzed in terms of the frequency of occurrence of the grammatical constructions that actually appear within the EGP.

14. References

Burston, Jack, and Maro Neophytou. 2014. "Lessons Learned in Designing and Implementing a Computer-adaptive Test for English." *The EUROCALL Review* 22(2):19-25. http://www.eurocall-languages.org/uploaded/ EUROCALL_Review/No22_2.pdf.

Burston, Jack, Maro Neophytou, and Iasonas Lamprianou. 2016. "Developing and Implementing a Computer-adaptive

Test for English: The SLUPE Experience." *CALL-EJ* 17(1):19-34. http://callej.org/journal/17-1/Burston2016.pdf.

Coste, Daniel, Janine Courtillon, Victor Ferenczi, Michel Martins-Baltar and Eliane Papo. 1987. *Un niveau seuil*. Paris: Didier.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Modern Languages Division, Strasbourg: Cambridge University Press.

Hawkins, John, and Luna Filipovic. 2011. *Criterial Features in L2 English: Specifying the Reference Levels of the Common European Framework. English Profile Studies*, *1*. Cambridge: UCLES/ Cambridge University Press.

Milton, James. 2010. "The Development of Vocabulary Breadth Across the CEFR Levels. A Common Basis for the

Elaboration of Language Syllabuses, Curriculum Guidelines, Examinations, and Textbooks Across Europe." *EUROSLA Monographs Series 1*, Communicative Proficiency and Linguistic Development, 211-232.

Van Ek, Jan, and John Trim. 1991. *Threshold 1990*. Strasbourg: Council of Europe.

—. 2001. *Vantage*. Cambridge: Cambridge University Press.

UCLES/CUP. 2011. *English Profile–Introducing the CEFR for English*. Cambridge: UCLES.