John M. Levis,
Tracey M. Derwing
and Murray J. Munro (eds.)

The Evolution of Pronunciation Teaching and Research

25 years of intelligibility, comprehensibility, and accentedness

BENJAMINS CURRENT TOPICS

121

The Evolution of Pronunciation Teaching and Research

Benjamins Current Topics

ISSN 1874-0081

Special issues of established journals tend to circulate within the orbit of the subscribers of those journals. For the Benjamins Current Topics series a number of special issues of various journals have been selected containing salient topics of research with the aim of finding new audiences for topically interesting material, bringing such material to a wider readership in book format.

For an overview of all books published in this series, please see
benjamins.com/catalog/bct

Volume 121

The Evolution of Pronunciation Teaching and Research
25 years of intelligibility, comprehensibility, and accentedness
Edited by John M. Levis, Tracey M. Derwing and Murray J. Munro

These materials were previously published in *Journal of Second Language Pronunciation* 6:3 (2020)

The Evolution of Pronunciation Teaching and Research

25 years of intelligibility, comprehensibility, and accentedness

Edited by

John M. Levis
Iowa State University

Tracey M. Derwing
Simon Fraser University & University of Alberta

Murray J. Munro
Simon Fraser University

John Benjamins Publishing Company
Amsterdam / Philadelphia

™ The paper used in this publication meets the minimum requirements of the American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

DOI 10.1075/bct.121

Cataloging-in-Publication Data available from Library of Congress:
LCCN 2022011334 (PRINT) / 2022011335 (E-BOOK)

ISBN 978 90 272 1137 8 (HB)
ISBN 978 90 272 5763 5 (E-BOOK)

John Benjamins Publishing Company · https://benjamins.com

Table of contents

FOREWORD

FOREWORD

Evolution of L2 pronunciation research and teaching[*]

25 years of intelligibility, comprehensibility, and accentedness

John Levis
Iowa State University

John Benjamins Publishing chose a special issue of the Journal of Second Language Pronunciation 6(3) (2020) as the content of this monograph. That special issue was first envisioned by John Levis, Editor of the Journal, who wanted to revisit what he viewed as an extraordinarily influential study for L2 pronunciation research and teaching. Murray Munro and Tracey Derwing's 1995 paper, "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners," published in the journal *Language Learning,* instigated tremendous changes in the research focus of second language pronunciation. (The paper was later reprinted as Munro & Derwing, 1999, again in *Language Learning.*) The authors provided evidence for three distinct, yet partially-related constructs: *intelligibility* (the degree to which a listener understands a speaker's intended message), *comprehensibility* (the degree of effort required for a listener to understand L2 speech) and *accentedness* (the degree of difference from an expected accent).

Since 1995, the three constructs have been invoked repeatedly in research and in relation to teaching. For example, they have been used to demonstrate the validity of functional load for prioritizing pronunciation segments (Munro & Derwing, 2006), validate the importance of listener judgments as measures of pronunciation improvement (Derwing, Munro & Wiebe, 1998), distinguish between accentedness and intelligibility goals for language teaching (Levis, 2005), make connections to fluency judgments (Derwing, Munro & Thomson, 2008), deconstruct the language features involved in comprehensibility judgments (Isaacs & Trofimovich, 2012), examine the effects of methodological choices on speech rating (O'Brien, 2016), relate judgments of comprehensibility to grammatical form (Ruivivar &

[*] An earlier version of this article was published as part of a special issue of the *Journal of Second Language Pronunciation* 6(3). https://doi.org/10.1075/jslp.20054.lev

https://doi.org/10.1075/bct.121.01lev

Collins, 2019), and help frame pedagogical goals emphasizing intelligibility (Levis, 2018).

The constructs have also been used to make connections of L2 pronunciation research to other areas of applied linguistics, especially in examining the role of pronunciation in language assessment (Isaacs & Harding, 2017), in showing the effects of instructional approaches (Foote & McDonough, 2017; Gordon & Darcy, 2016), in measuring pronunciation development in workplace and classroom contexts (Derwing, Munro, Foote, Waugh & Fleming, 2014; Nagle, 2017), in demonstrating differences in understanding World Englishes (Kang, Thomson, & Moran, 2018), and in showing connections between L2 pronunciation and social attitudes (Reid, Trofimovich & O'Brien, 2019). In addition, the constructs have been successfully applied in the study of longitudinal naturalistic L2 acquisition (e.g., Derwing & Munro, 2013).

These studies and many more demonstrate the continuing influence and flexibility of the original insights of the 1995 study. Although intelligibility had previously been studied widely for L1 listeners, extensive research into L2 intelligibility needed a way to distinguish lack of understanding from listener challenges in processing speech, clearly operationalized constructs, and evidence that intelligibility and accentedness were not closely related. In addition, Munro and Derwing clarified definitions for the three terms, a step critical for a productive research agenda. The impact of this seminal work is due to the care with which the three constructs were defined, measured, and shown to be distinct. In this volume, Munro and Derwing annotate their original research looking at the research since 1995. They also reconsider elements of the original paper that have been neglected or misunderstood and provide new analyses of the original findings.

This monograph includes several categories of papers. First, it includes not only a re-analysis and commentary of the original 1995/1999 paper (making it, possibly, the first paper to have been published four times), it also includes a reconsideration of another oft-cited paper about the Intelligibility and Nativeness Principles (Levis, 2005). This updated paper argues that the two principles are relevant to L2 pronunciation for any language, and makes a case for the superiority of the Intelligibility Principle while calling for approaches based on the Nativeness Principle to be consigned to the past.

The monograph also contains two instructional studies with long-term results. Beth Zielinski and Elizabeth Pryor examine the individual trajectories of beginner and intermediate immigrant learners in Australia. The groups differed in their amount of English use over time, with the intermediate learners using more English, but individual variation demonstrated that otherwise similar groups showed a wide range of comprehensibility outcomes.

In another innovative study, Leif French, Nancy Gagné and Laura Collins look at the effects of a five-month intensive English course on French-speaking high school students' comprehensibility, fluency and accentedness. Four years after the course, the students who took part in the intensive course had significantly more positive ratings for comprehensibility and fluency than the students who did not. Accentedness ratings for the two groups, however, did not differ.

This volume also highlights two studies of ratings of second language speech, with a special emphasis on applications to other languages and factors that affect differences in how the constructs are measured. In an extension of the research on English to other languages, Charles Nagle and Amanda Huensch replicate Munro and Derwing's 1995 study for L2 Spanish. They demonstrate again that the three constructs are partially independent, and that accentedness is only loosely related to the other two constructs, while comprehensibility and intelligibility are more closely related to each other.

Talia Isaacs and Ron Thomson compared the ratings of experienced teachers and novice raters for Mandarin and Slavic language speakers, connecting their ratings with measures of prosody, segments and temporal features of speech. They also asked raters to describe why they rated as they did. Results showed that both suprasegmental and segmental deviations were related to ratings. Experienced teachers also provided longer reports about why they rated as they did, indicating that having a way to talk about language results in more informative comments.

Two chapters also address L2 pronunciation's connections with NNS-NNS speech. Pavel Trofimovich, Charles Nagle, Mary O'Brien, Kym Taylor Reid, Sara Kennedy, and Lauren Strachan looked at how comprehensibility rating differs as a function of interaction and task. L2 English university students from different language backgrounds took part in three collaborative and interactive tasks. They rated their partner's comprehensibility every two or three minutes. Rather than remaining static, mutual comprehensibility went from high to low and then increased to high again by the end of the three tasks. The authors argue that changes in comprehensibility in L2-L2 interactions are normal and that there is room for teasing apart the effects of interaction, task, and time on comprehensibility measurements.

In another welcome chapter, Veronika Thir tests an assertion of the Lingua Franca Core (Jenkins, 2000) that the NURSE vowel, the only vowel quality feature in the LFC, is essential for international intelligibility. Her findings cast doubt on the importance of the NURSE vowel, but provide evidence for the importance of the TRAP-DRESS contrast. She interprets the findings in light of their relative functional load. This study renews attention to LFC pronunciation research by connecting it to other key areas of pronunciation research.

Finally, Charlotte Vaughn and Aubrey Whitty connect comprehensibility to social evaluations of speech. It is undeniable that speech can be intelligible and comprehensible yet still evoke prejudicial judgments. This study tests the processing fluency hypothesis by examining social evaluations of Korean L2 English speakers under two conditions: when a written text of their speech is provided, or when it is withheld. The study found that social evaluations were downgraded when orthography was first provided then withheld, suggesting that modest changes to context (i.e., providing a written text) can affect listener views of comprehensibility and social judgments.

References

Derwing, T. M., & Munro, M. J. (2013). The development of L2 oral language skills in two L1 groups: A 7-year study. *Language Learning*, 63(2), 163–185. https://doi.org/10.1111/lang.12000

Derwing, T. M., Munro, M. J., Foote, J. A., Waugh, E., & Fleming, J. (2014). Opening the window on comprehensible pronunciation after 19 years: A workplace training study. *Language Learning*, 64(3), 526–548. https://doi.org/10.1111/lang.12053

Derwing, T. M., Munro, M. J., & Thomson, R. I. (2008). A longitudinal study of ESL learners' fluency and comprehensibility development. *Applied Linguistics*, 29(3), 359–380. https://doi.org/10.1093/applin/amm041

Derwing, T. M., Munro, M. J., & Wiebe, G. (1998). Evidence in favor of a broad framework for pronunciation instruction. *Language Learning*, 48(3), 393–410. https://doi.org/10.1111/0023-8333.00047

Foote, J. A., & McDonough, K. (2017). Using shadowing with mobile technology to improve L2 pronunciation. *Journal of Second Language Pronunciation*, 3(1), 34–56. https://doi.org/10.1075/jslp.3.1.02foo

Gordon, J., & Darcy, I. (2016). The development of comprehensible speech in L2 learners: A classroom study on the effects of short-term pronunciation instruction. *Journal of Second Language Pronunciation*, 2(1), 56–92. https://doi.org/10.1075/jslp.2.1.03gor

Isaacs, T., & Harding, L. (2017). Pronunciation assessment. *Language Teaching*, 50(3), 347–366. https://doi.org/10.1017/S0261444817000118

Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition*, 34(3), 475–505. https://doi.org/10.1017/S0272263112000150

Jenkins, J. (2000). *The phonology of English as an international language*. Oxford: Oxford University Press.

Kang, O., Thomson, R. I., & Moran, M. (2018). Empirical approaches to measuring the intelligibility of different varieties of English in predicting listener comprehension. *Language Learning*, 68(1), 115–146. https://doi.org/10.1111/lang.12270

Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly*, 39(3), 369–377. https://doi.org/10.2307/3588485

Levis, J. M. (2018). *Intelligibility, oral communication, and the teaching of pronunciation.* Cambridge University Press. https://doi.org/10.1017/9781108241564

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, 45(1), 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Munro, M. J., & Derwing, T. M. (1999). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Special reprint of the decade's best papers in Language Learning*, 49, 285–310. https://doi.org/10.1111/0023-8333.49.s1.8

Munro, M. J., & Derwing, T. M. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System*, 34(4), 520–531. https://doi.org/10.1016/j.system.2006.09.004

Nagle, C. L. (2017). Individual developmental trajectories in the L2 acquisition of Spanish spirantization. *Journal of Second Language Pronunciation*, 3(2), 218–241. https://doi.org/10.1075/jslp.3.2.03nag

O'Brien, M. G. (2016). Methodological choices in rating speech samples. *Studies in Second Language Acquisition*, 38(3), 587–605. https://doi.org/10.1017/S0272263115000418

Reid, K. T., Trofimovich, P., & O'Brien, M. G. (2019). Social attitudes and speech ratings: effects of positive and negative bias on multi-age listeners' judgments of second language speech. *Studies in Second Language Acquisition*, 41(2), 419–442. https://doi.org/10.1017/S0272263118000244

Ruivivar, J., & Collins, L. (2019). Nonnative accent and the perceived grammaticality of spoken grammar forms. *Journal of Second Language Pronunciation*, 5(2), 269–293. https://doi.org/10.1075/jslp.17039.rui

Foreign accent, comprehensibility and intelligibility, redux*

Murray J. Munro[1] and Tracey M. Derwing[1,2]
[1] Simon Fraser University | [2] University of Alberta

We revisit Munro and Derwing (1995a), providing retrospective commentary on our original methods and findings. Using what are now well-established assessment techniques, the study examined the interrelationships among accentedness, comprehensibility, and intelligibility in the speech of second-language learners. The key finding was that the dimensions at issue are related, but partially independent. Of particular note was our observation that speech can be heavily accented but highly intelligible. To provide a fresh perspective on the original data we report a few new analyses, including more up-to-date statistical modeling. Throughout the original text we intersperse insights we have gained since the appearance of the 1995 paper. We conclude with retrospective interpretations, including thoughts on the relevance of the study to contemporary second language teaching and especially pronunciation instruction.

Keywords: accentedness, intelligibility, comprehensibility, pronunciation

In 2019, when John Levis told us of his plan to commemorate the 25th anniversary of Munro and Derwing (1995a), we were honoured and humbled. For two relatively new scholars, this study represented a lot of work, but we were both excited to do it. We planned it in Birmingham, Alabama, in April of 1993. The year before, we had conducted our first accent study, and we were brimming with questions and ideas for new investigations. We had already collaborated on several other projects, but this particular study propelled us into a career-long partnership that has been rich and satisfying. We still have a lot of questions but many more people are working on them now, and our field is growing to an extent we couldn't

* An earlier version of this article was published as part of a special issue of the *Journal of Second Language Pronunciation* 6(3). https://doi.org/10.1075/jslp.20038.mun

https://doi.org/10.1075/bct.121.02mun

have imagined in 1993. Much of that growth can be attributed to the efforts of John Levis, who started the Pronunciation in Second Language Learning conference (PSLLT), its proceedings, and the *Journal of Second Language Pronunciation*, among his many other undertakings.

Rather than simply reprint the original paper, we have shortened several sections and have used a different font and colour for the original parts of the paper (blue Myriad Pro (sans serif) is old; new is in regular type). We comment on what we did 25 years ago, adding information about relevant studies since, and making suggestions for future research. The passive voice has been employed here and there. (The copy-editor at *Language Learning* forbade it in the original paper, but we find it quite useful.) Ellipses indicate omissions of text from the original article. Also, after considerable searching through archives, we located some of our original data, which we have used to run a few new analyses to provide some slightly new perspectives on our findings.

The paper began as follows: For several decades, pronunciation experts have stressed improved intelligibility as the most important goal of pronunciation teaching. As early as 1949, Abercrombie argued that most "language learners need no more than a comfortably intelligible pronunciation" (p. 120).... However, up to the present time, there has been a heavy emphasis in classrooms on accent reduction, with native like pronunciation as the target.

Numerous studies have shown that native-speaker (NS) listeners tend to downgrade nonnative speakers (NNSs) simply because of foreign accent (e.g., Brennan & Brennan, 1981…). Thus, second language instructors, curriculum designers, and writers of textbooks may feel obliged to focus attention on accent reduction, without regard to specific features that may interfere with intelligibility, because any accentedness is seen as a problem.… However, there is as yet no indication that reduction of accent necessarily entails increased intelligibility. The effects of nonnative-like pronunciations on intelligibility are far from clear. In the present study, we have attempted to gain a better understanding of the interrelationships among accentedness, intelligibility, and listeners' perceptions of accent and of comprehensibility.

In 1995, we reviewed studies of error gravity hierarchies; that is, rankings of linguistic errors according to their impact on intelligibility. L2 phonology, grammar, vocabulary, fluency, discourse organization and overall error frequency were identified as causing problems for listeners, but there was little convergence of findings. The apparent contradictions in all of these studies may be at least partially explained by the differences in the target languages under study, as well as by differences in methodology (cf. Schairer, 1992). The effects of second language accent on intelligibility remain unresolved.

We also reviewed accent gravity hierarchies and came to the conclusion that [n]ot only is there little empirical evidence regarding the role of pronunciation in determining intelligibility, but there is no clear indication as to which specific aspects of pronunciation are most crucial for intelligibility. Several researchers have found evidence that prosodic errors are more serious than segmental errors (Anderson-Hsieh, Johnson, & Koehler, 1992…). On the other hand,… Fayer and Krasinski (1987) argued that segmental errors are more detrimental to comprehension.

Intelligibility, comprehensibility and pronunciation

To gain a better understanding of these issues, the relationship between foreign accent and speech intelligibility must be examined. Intelligibility may be broadly defined as the extent to which a speaker's message is actually understood by a listener, but there is no universally accepted way of assessing it.

… In this study, we chose to obtain two types of assessments of listener comprehension in addition to foreign accent ratings. First, we adopted a measurement of intelligibility using a technique similar to that used by Gass and Varonis (1984), i.e., transcriptions made by listeners.… Second, we asked listeners to assign [perceived] comprehensibility judgments using a 9-point Likert scale. We then examined the relationships between these scores and their relationship with global foreign accent scores. Note that the expression "perceived comprehensibility" was not our preference for the 1995 paper. Rather, "perceived" was added during the review process at the insistence of a reviewer. We now see that addition as a mistake, though we have left the word intact in the original text that follows. We strongly advise against such usage and comment further on the issue in our retrospective interpretations near the end of this paper.

On the basis of previous work, we anticipated that intelligibility, perceived comprehensibility, and accentedness would be correlated. Here we must point out two excellent studies that greatly influenced the constructs we chose to investigate. Varonis and Gass (1982) is the first L2 study we know of to use a comprehensibility scale similar to the one we chose. Their listeners rated utterances on a 5-point comprehensibility continuum ranging from "I understood this sentence easily" to "I didn't understand this sentence at all." [They] argued that the "main factor involved in judgments of pronunciation was overall comprehensibility or ease of interpretation" (p. 127). However, it cannot be concluded, even when content is controlled, that accent and intelligibility [or accent and comprehensibility] are identical dimensions; that is, the focus of listeners' perception of accent may be somewhat different from the focus of a judgment of comprehensibility. In the second influential paper, Gass and Varonis (1984) articulated their clear understanding that comprehensibility and intelligibility are distinct. They used a sentence transcription task, explaining that they "were interested in how much was understood, rather than just intuitive judgments of ease of comprehensibility." (p. 68). However, throughout that paper, they used "comprehensibility" to refer to what we now call "intelligibility," even though their 1982 work had operationalized "comprehensibility" in an entirely different way. The authors can't be blamed for these differences in usage because the pronunciation field itself suffered from rampant labelling inconsistencies, and we were at times stymied by contradictions and conflations in the literature. However, we hoped, in our work from 1995 on, to "nail down" an empirically-based three-way distinction and to encourage movement toward a standard terminology.

Method

Speech materials

Speakers

The speech samples used in this experiment were elicited from 10 native speakers of Mandarin (5 male and 5 female), who had learned English after puberty. All were proficient speakers of English who had scored no less than 550 on the TOEFL, and all had spent a minimum of one year in Canada as graduate students at the University of Alberta. Assessments by the authors, both of whom have had many years of experience with English as a second language (ESL) students, indicated that their English pronunciation ranged from moderately to heavily foreign-accented. Recordings were also made of 2 native speakers of Canadian English (1 male and 1 female).

Recording

Individual recording sessions were held in a sound-treated room with high fidelity audio equipment. We gave the speakers a page of cartoons that illustrated an amusing story and asked each person to describe the events depicted. No preparation was allowed; nor were there any verbal exchanges between the experimenter and the speaker during the narration. The entire task took two to three minutes for each participant. To simplify the stimulus preparation procedure, we digitally rerecorded the speech samples at 10 kHz using a Kay Computerized Speech Lab (CSL). We used the waveform editing feature of the CSL to divide the speech samples into shorter excerpts that were of sufficiently short duration to be transcribed by listeners after a single listening. We selected three excerpts from the initial 30 seconds of the narrative from each speaker, for a total of 36 samples. It was not practical to attempt to break the original recordings down into new samples of exactly identical durations, because this would have resulted in utterances that did not necessarily begin or end at phrasal or clausal boundaries. Instead, the excerpts ended at locations of natural pauses in the utterances, as identified by us. As a result, the final stimulus set of 36 samples varied somewhat in length: the mean length was 10.7 words, with a range of 4 to 17 words. We rerecorded the stimuli in random order onto a cassette tape.

Listeners

The listeners were 18 native speakers of English who were enrolled in either an introductory linguistics course or an ESL teaching methodology course at the University of Alberta. All reported normal hearing, and all had a basic knowledge of articulatory phonetics. We paid each person an honorarium of $10 upon completion of the experiment.

The cartoon story, referred to in our subsequent work (e.g., Derwing & Munro 1997) as the "Hunting Story," was taken from a Canadian secondary school French textbook (Rondeau, 1972). It depicts two men who leave home on a deer-hunting trip, only to be foiled by a rainstorm. When the sun eventually comes out, they find themselves taking photographs of the deer instead of shooting them. We viewed the cartoon as a "feel-good" story; some of our reviewers, however, accused us of potentially traumatizing our participants by showing them depictions of rifles and hunting, which led us to develop the ubiquitous suitcase story featured in a number of later studies (e.g., Derwing, Munro, Foote, Waugh & Fleming, 2014; Derwing, Munro & Thomson, 2008; Derwing, Rossiter, Munro & Thomson, 2004; French, Gagné & Collins, this volume; Isaacs & Thomson, this volume).

Although we used high quality recording equipment, the technology available to most researchers at the time was much less well-developed than it is today. We followed the accepted, time-consuming approach of first making analog recordings on audio tape and then digitizing them for editing and analysis, in this case at an acceptable resolution of 10 bits. Today, 16-bits (CD quality) or better is the norm. Because we collected listener ratings in free-field conditions with multiple listeners, we could not present the stimuli digitally. Instead, we had to re-record the digitized excerpts on tape for presentation.

Procedure

We held two listening sessions. During Session 1, we handed the listeners booklets with numbered spaces for transcriptions of each of the 36 utterances. Each space in the booklet also included a Likert scale numbered from 1 to 9. In previous work (Munro & Derwing, 1994) we found a scale of this size to be effective for eliciting judgments of nonnative speech. We instructed the listeners to listen carefully to each utterance and then write out in standard orthography exactly what they had heard; in other words, to write the utterances word for word. (This was the intelligibility task.) Upon completion of each orthographic transcription, they assigned a perceived comprehensibility rating by circling a number from 1 to 9, where 1=extremely easy to understand and 9=impossible to understand.

We presented the stimuli through a high fidelity playback system in a quiet room. Before beginning the task, we provided the listeners with two practice stimuli for orthographic transcription and rating. During the experiment, one of the experimenters controlled the tape by pressing a pause button at the end of each utterance. A new stimulus was not presented until all listeners had finished transcribing the previous one. The entire session lasted approximately 20 minutes.

Session 2 was held four days later. This time we presented the listeners with the same 36 stimuli, but we asked them to rate the degree of foreign (non-English) accent in each sample. We again used a 9-point scale, where 1=no foreign accent and 9=very strong foreign accent. The same two example stimuli were provided for practice at the beginning of the session. The session lasted approximately 10 minutes.

In an ideal situation, we would have carried out individual listening sessions with headphones in a sound-treated lab via computer-based presentation and response software. But in this exploratory study we wished to obtain a substantial data set as quickly and efficiently as possible. We therefore opted for group data collection in a quiet classroom. In James Flege's lab, MS-DOS presentation software had been developed for rating tasks, but it would have been too unwieldy for us (especially as Macintosh devotees) to use it "live" in a group task. Like all computer users of the day, we were deeply fearful of system crashes, often referred to as the "blue screen of death." Instead, Murray coded a Macintosh-based custom software package for designing, presenting, and re-recording stimuli, which we used in nearly all of our studies (whether individual- or group-based) in the 1990s and the early 2000s.

Despite our extensive experience with individual rating sessions in a lab context, we faced some challenging unknowns:

Length and number of sessions

We were wary of demanding too much of listeners in what would surely be a boring task. However, we had little sense of how the listeners would feel about transcribing 36 utterances and assigning two types of ratings to each one. Our decision to divide the work into two sessions arose from this concern, along with our uncertainty about how the responses on one part of the task might affect the other judgments. In fact, subsequent work by O'Brien (2016) yielded no effect of ordering on L2 speech ratings and indicated that multiple judgments can be reliably collected in a single session.

Size of scale

Of all the methodological considerations, this one has perhaps engendered the most controversy. Debate still exists about optimal scale size. Using too small a scale might obscure perceivable differences between speakers or speech samples, but too large a scale might be unmanageable for raters and might compromise accuracy and reliability. Listeners in our earlier study of accentedness (Munro & Derwing, 1994) had found a 9-point scale easy to use. For comprehensibility, however, we had nothing to go on. Our intent was to compare the two sets of ratings directly, and our colleague Helen Southwood, an expert in scaling, had encouraged us to use at least 9 points for both scales. We comment further about scales later in this paper.

Results

Coding

We transcribed the complete set of utterances in broad phonetic transcription. We played the stimulus items as many times as necessary, so that the total numbers of phonemic, phonetic, and grammatical errors could be tallied.…

We defined phonemic errors as either the deletion or insertion of a segment, or the substitution of a segment that was clearly interpretable as an English phoneme different from the correct one. The total phonemic errors for the Mandarin speakers' productions ranged from 0 to 3 per utterance, with a mean of 0.9. Phonetic errors involved the production of a segment in such a way that the intended category could be recognized but the segment sounded noticeably nonnative. The number of phonetic errors per utterance ranged from 0 to 4, with a mean of 1.6.

We found only 19 morphosyntactic errors in the entire set: 1 utterance contained 3 errors, 3 utterances contained 2 errors and 10 utterances contained 1 error. Of the 30 speech samples produced by the nonnative speakers, 15 were error free. Of the grammatical errors identified, 6 involved inappropriate use of prepositions, 3 involved errors in subject-verb agreement, 3 were errors in verb tense and 2 were errors in verb form. We also noted one instance of each of the following types of errors: inappropriate article, incorrect number, missing subject, missing object, and missing relative pronoun.

We rated the intonation of each speech sample independently on a scale where 1=native-like and 9=not at all native-like. We then compared the ratings. In any cases where we had assigned ratings more than one scalar unit apart, we played the stimuli again and re-evaluated those stimuli independently. In four cases the final ratings were two scalar units apart, all others were identical or only one scalar unit apart. We averaged the ratings for the final analyses. The scores ranged from 1.0 to 8.5, with a mean of 3.8.

We recognized that making an assessment of prosody independent of other aspects of the speech was difficult, and that there was no guarantee that our judgments would be unaffected by factors other than prosody. In later research, we aimed to circumvent this problem by using low-pass filtered utterances. In this approach, inspired by Van Els and de Bot (1987), acoustic components above 300 Hz were removed (225 Hz for male voices), leaving the speech largely unintelligible, but preserving most of the intonational information. (Derwing & Munro, 1997; Munro, 1995)

We coded the transcriptions provided by the listeners for exact word matches, substitutions (defined as the substitution of one word for a phonetically and semantically similar word, e.g., *who* for *he),* novel words (defined as the insertion of a word bearing no phonological resemblance to a word in the stimulus utterance), and regularizations (e.g., *he walks* for *he walk*). We also identified word omissions and categorized them as either content (nouns, verbs, adjectives, adverbs) or function words (particles, determiners).

In a subsequent study (Derwing & Munro, 1997) we coded transcriptions according to the criteria above, but we also coded errors as *trivial* and *nontrivial*. Trivial errors were those where intelligibility was clearly not at risk, for example, regularizations where the listener corrected a plural of ‘two mans’ to ‘two men’ or left out a repeated word, transcribing ‘the car’ instead of ‘the, the car’, thus reducing the penalty to the speaker that these errors levied. Transcriptions, nonetheless, are not a perfect measure, partly because word recognition is not the same as comprehension (Zielinski, 2008). In fact, no measure of intelligibility is ideal (Derwing & Munro, 2015). Transcription errors can occur when the listener’s mind wanders and when memory fails. The same limitations apply to other intelligibility assessments such as true/false sentence verifications, summaries of minilectures, and responses to comprehension questions. De Weers (2020), recently implemented a very promising, easy-to-administer technique, that is closer to online than other tasks, rapid, and engaging. She asked listeners to immediately repeat utterances produced by L2 speakers and later transcribed the recorded repetitions for analysis.

Analyses

Judgment tasks

We tabulated the comprehensibility and accentedness judgments for each stimulus. The mean perceived comprehensibility ratings (pooled across listeners) ranged from 1.0 to 7.6. As expected, the six samples produced by the native speakers of English received the six lowest mean accent scores (i.e., the most native-like ratings). In addition, five of the native English samples received the best mean perceived comprehensibility scores (ranging from 1.0 to 1.4). However, one of the native English samples was rated worse (2.4) than 11 of the nonnative samples. Although this stimulus received a rating indicating that it was largely heard as unaccented (viz. 1.6), for some reason it was rated as less comprehensible than many of the other samples. This finding does not seem surprising, given that even nonpathological native speech may vary in comprehensibility because of such factors as rate of speech, speech clarity, voice quality, and word choice.

Figures 1, 2 and 3 illustrate the distributions of the accent, perceived comprehensibility, and intelligibility scores for the stimuli produced by the native Mandarin speakers. The accent ratings (Figure 1) are fairly evenly distributed across Categories 2 to 8. Only a very small number of the judgments (4%) were ratings of 1, indicating no foreign accent. The listeners were apparently quite successful at recognizing which speech samples were produced by the nonnative speakers. The comprehensibility judgments (Figure 2) show a strikingly different pattern. Twenty-two percent of the samples were rated as extremely easy to understand (Category 1) and 64% of the ratings were in Categories 1, 2, or 3. The skewed distribution indicates that the perceived comprehensibility ratings were, on the whole, less harsh than the accent ratings.

To provide a new perspective on the same data we created two new frequency plots (Figure A, 2020). These display the proportion of rating responses on each of the 9-point scales separately according to intelligibility. For instance, the striped bar for ‘3’ corresponds to the proportion of instances in which a particular listener assigned a rating of 3 to an utterance that the same listener found 100% intelligible, while the solid bar represents the parallel proportion of ratings when intelligibility was less than 100%. For comprehensibility, both sets of

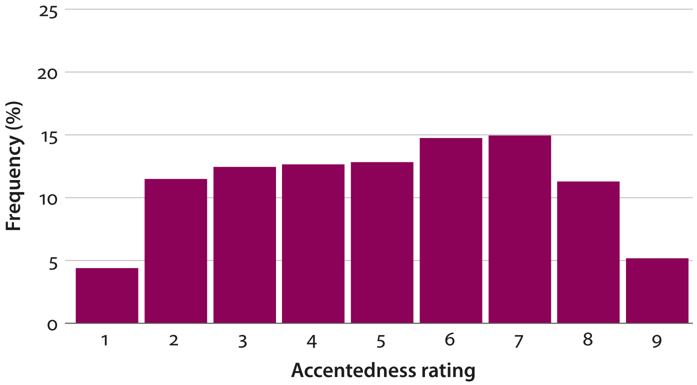


Figure 1. Distribution of listener ratings of the strength of foreign accent (1 = no foreign accent; 9 = very strong foreign accent). [Recreated]

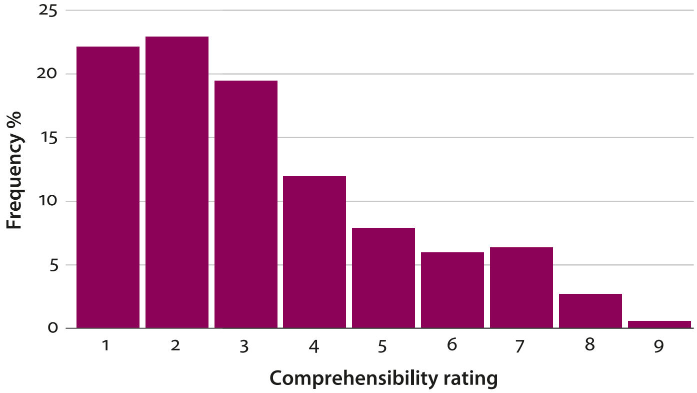


Figure 2. Distribution of listener ratings of perceived comprehensibility (1 = extremely easy to understand; 9 = impossible to understand). [Recreated]

scores show positive (rightward) skew, but the mode is at 1 for the 100% items, which cover nearly a third of all the ratings. Only about 10% of ratings were '1' for the imperfectly transcribed stimuli, and the mode fell at 3. These scores show greater skew overall than the scores for the perfectly transcribed utterances. In contrast, the two sets of accentedness ratings show quite similar distributions: in both cases, we see relatively flat distributions with between 10 and 15% of ratings at each scale point from 3 to 7.

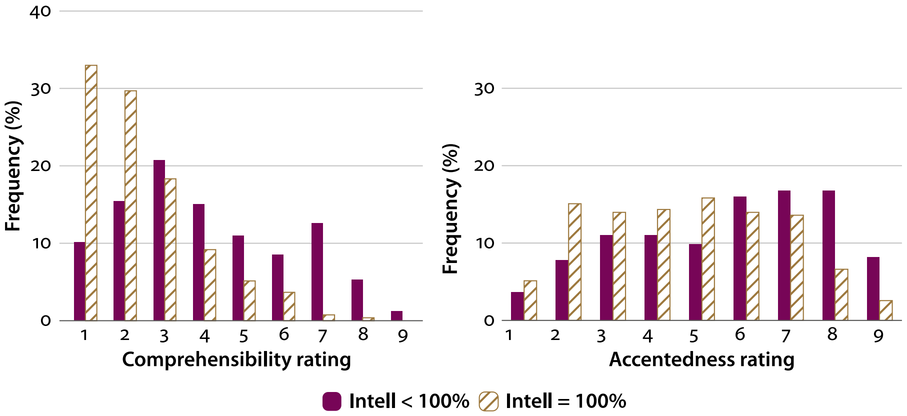


Figure A (2020). Distributions of comprehensibility (left) and accentedness (right) ratings for items with two corresponding categories of intelligibility (< 100%; = 100%). Higher ratings indicate worse performance

Orthographic transcription task

The frequencies of the various types of transcription errors appear in Table 1. The orthographic transcriptions of the native English speakers' productions were not completely free of errors; in fact, 44 errors were noted, most of which we classified as substitutions. The number of errors in the transcriptions of the Mandarin speakers was much higher (636), but it must be remembered that there were 10 Mandarin speakers and only 2 native English speakers. When this difference is taken into account, it can be seen that the mean number of errors per speaker was nearly three times greater in the transcriptions of the Mandarin speakers than in those of the native English speakers (63.6 vs. 22.0).

Table 1. Frequency of transcription error types [Recreated]

	Mandarin Speakers		English Speakers	
Error Type	**Count**	**%**	**Count**	**%**
Omission (Function Word)	135	21	11	25
Omission (Content Word)	154	24	9	20
Novel Word	76	12	1	2
Substitution	183	29	22	50
Regularization	88	14	1	2
Total	**636**		**44**	
Mean # of Errors per Speaker	63.6		22	

We assigned each of the 648 orthographic transcriptions an intelligibility score on the basis of the number of words that exactly matched our corresponding transcription. We also computed an overall intelligibility score for each of the 36 utterances by taking the mean of the 18 listeners' scores for the utterance. The scores for the Mandarin speakers' productions ranged from 39% to 100%; the native English speakers' production scores ranged from 94% to 99%. Five productions were 100% intelligible to all listeners. Surprisingly, these utterances were all produced by Mandarin speakers. The listeners' success in transcribing these stimuli was probably not due to [short] utterance length, because the lengths varied from 7 to 13 words. These items were therefore representative of stimuli in the middle of the length range. An additional seven stimuli from the Mandarin speakers were transcribed with intelligibility scores equal to or above that of the native English stimulus with the lowest intelligibility score. Finally, five nonnative stimuli were transcribed with at least one error by every listener.

Figure 3 illustrates the distribution of intelligibility scores for the stimuli produced by the Mandarin speakers. Again, the distribution is highly skewed; the largest category by far (64%) is the one including scores from 91% to 100%. In fact, 53% (275) of the transcriptions of the nonnative stimuli received accuracy scores of 100%. Moreover, of the orthographic transcription errors reported in Table 1, more than one third were trivial errors: either omissions of function words or regularizations. On the whole, then, it appears that the nonnative speech samples used in this study were highly intelligible. The distribution of these scores resembles the distribution of the perceived comprehensibility scores (Figure 2) more closely than that of the foreign accent scores (Figure 3), though it differs in some respects from both.

We excluded one nonnative stimulus item from further analyses on a number of grounds. First, it received an overall intelligibility score that was considerably lower (39%) than that of the next worst stimulus (68%). Second, it was the first item heard by the listeners (after the practice items) and third, it was relatively long (15 words). Possibly the listeners were not prepared for a stimulus of this level of difficulty at the outset of the task. Thus, their poor performance on this item may reflect something other than poor comprehension.

The issue raised above illustrates one of the drawbacks of using a particular stimulus randomization more than once, a problem inherent in group rating tasks. Although best practice is to use a different randomization for each listener, we adopted a compromise approach in later group tasks by assigning listeners to small groups, each of which heard a different randomization. For the record, we know of no important differences between the results of our group tasks and those

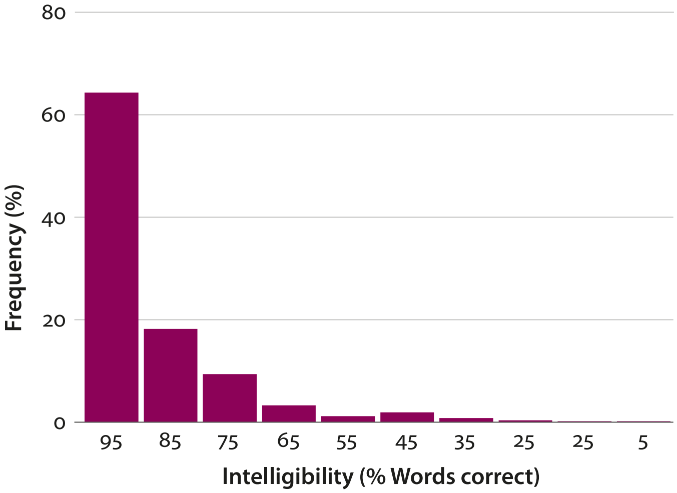


Figure 3. Distribution of intelligibility scores (percentages of words transcribed correctly per utterance). [Recreated]

we obtained in our (numerous) lab-based tasks involving a unique randomization for each listener.

A second issue of note here is the fact that prior to the rating task, the listeners knew nothing of the content of the narratives that they would hear. As a result, their familiarity with the story would increase over the course of the task. Early-encountered items might therefore be judged differently from later ones. In subsequent work, we initiated the rating sessions by showing the listeners the cartoon story, thus avoiding a familiarity effect.

Cross-task comparisons

One issue here was whether there were significant inter-listener differences in the patterns of ratings under the two rating conditions (accent and perceived comprehensibility). First, we assessed interrater reliability on the two ratings tasks by computing intraclass correlations (Shrout & Fleiss, 1979). The correlations were very high for both the comprehensibility ratings (0.96, $p < 0.05$) and the accent ratings (0.98, $p < .05$), indicating that the raters tended to agree with one another on both.

Subsequent studies in our labs and those of our colleagues have shown that 9-point Likert-type judgments yield highly reliable results for both dimensions with intraclass correlations typically exceeding .9. Similar levels of reliability have been reported for quasi-continuous scaling, in which the raters see no labelled points, and the scale has 100 or even 1024 levels (see Munro, 2018). Some scholars have raised the issue of rater bias (e.g., Lindemann & Subtirelu, 2013). We acknowledge that all rating data are subject to a wide range of biases and that reactions to L2 speech in the real world can be influenced by discriminatory attitudes. This is a serious social issue, but in intelligibility and comprehensibility research, the impact of many types of bias can be minimized. For instance, holding L1 constant in a rating task makes it largely immaterial whether or not a particular listener is biased against the accent at issue. While we might expect a biased rater to assign harsher ratings overall, extensive reliability data indicate that speakers tend to be ranked in much the same order by raters, irrespective of "harshness." This is why our research focuses on the relative rankings of speech samples. We are interested in how one speaker compares to another on the dimensions we explore, not on whether one rater is stricter than another.

....

We calculated Pearson correlation coefficients *(r)* for each listener for the accent and intelligibility judgments of the 29 nonnative speech samples and the total numbers of phonemic, phonetic, and grammar errors, intonation ratings and utterance length (in words).... In a footnote we observed that [i]nclusion of the ratings of the native speaker samples might have led to spuriously high correlations, because all but one of these samples received very good ratings on both scales. We strongly discourage other researchers from including ratings of native speech in their calculations.

We first assessed the relationships among the three data sets obtained from the listeners. For all but 1 of the 18 listeners there was a significant positive correlation between the perceived comprehensibility ratings and the accent ratings: evidence that perceived comprehensibility and accent were nonorthogonal dimensions for most listeners. However, the significant correlations ranged from 0.41 to 0.82, indicating that the strength of the relationship between perceived comprehensibility and accent varied a great deal from listener to listener. For 15 of the listeners (83%) there was a significant negative correlation between the perceived comprehensibility (high to low) and transcription intelligibility scores (low to high). The relationship between these two variables suggests that the listeners' perceived comprehensibility ratings tended to reflect their actual understanding of the utterances, measured by their ability to write down exactly what they had heard (not entirely surprising, given that the judgments were made immediately after the transcription task). Again, however, the significant correlations showed a wide range (–0.44 to –.0.90). Finally, for only 5 listeners (28%) was there a significant correlation between the accent scores and the orthographic transcription (intelligibility) scores. These values ranged from –0.37 to –0.48.

Next, we examined two subsets of stimuli. First, we considered only the five stimuli that were transcribed perfectly by all 18 listeners. Figures 4 and 5 show the distributions of comprehensibility and accent ratings for these stimuli. As expected, the perceived comprehensibility scores tended to be quite low (indicating that the stimuli were easy to understand): as a result, the distribution in Figure 4 is highly skewed. In contrast, Figure 5 illustrates that the accentedness judgments are much more evenly distributed across the range of possible scores. That none of the utterances ever received a foreign accent rating of 1 (perfectly native-like) indicates that all listeners believed that they were produced by non-native speakers of English. Furthermore, the listeners apparently perceived a wide range of accentedness in stimuli that were nonetheless perfectly transcribed. Highly intelligible stimuli were not necessarily assigned low accent scores.

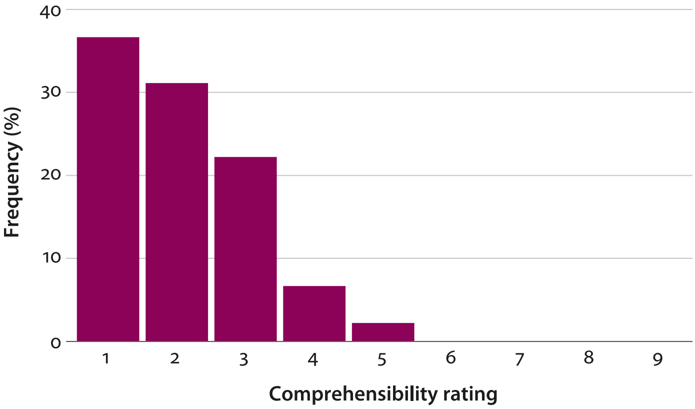


Figure 4. Distribution of comprehensibility scores for the five stimuli transcribed orthographically without error by all listeners. [Recreated]

Table 2 gives the numbers and percentages of the significant correlations between the various stimulus assessments and the three sets of listener scores. The majority of listeners (over 70% in all cases) showed significant correlations between the phonemic, phonetic, intonation, and grammar scores and the accent scores. This finding suggests that our assessments do indeed reflect stimulus properties that the listeners took into account when making their accent judgments. The numbers of listeners showing correlations between these properties and the perceived comprehensibility scores were somewhat lower, however. This tendency was particularly true for the two categories of segmental errors; only 44% and 11% of the listeners showed correlations with the phonemic and the phonetic scores, respectively. This finding suggests that these stimulus properties have more relevance to perceptions of accent than to perceptions of comprehensibility. Further support for this hypothesis surfaced when we considered their relationships with

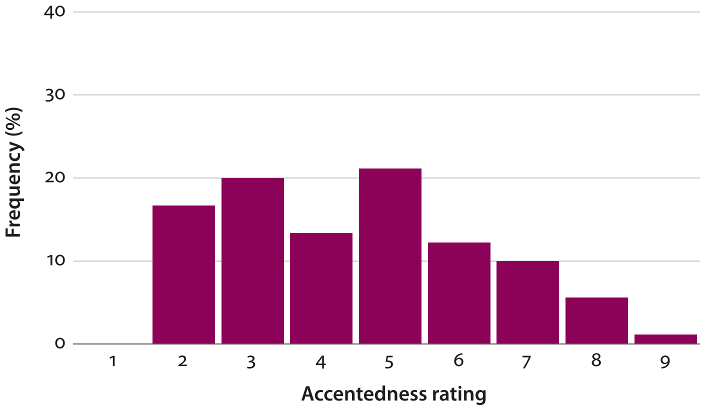


Figure 5. Distribution of foreign accent scores for the five stimuli transcribed orthographically without error by all listeners. [Recreated]

the orthographic transcription intelligibility scores. Only a handful of listeners showed relationships between any of the stimulus properties and the intelligibility scores. In fact, none showed such a relationship for the phonetic scores. Finally, utterance length did not correlate with any of the scores. Apparently, the stimuli were of suitable length for the listeners to make the required judgments and perform the orthographic transcription task. Had some of the utterances been too long, we would have expected some significant correlations with utterance length.

Table 2. Number of significant correlations between perceived comprehensibility, accent, and intelligibility
Scores and a stimulus measure ($p < .05$) [Recreated]

	Comprehensibility		Accent		Intelligibility (Words Correct)	
Stimulus Measure	**Count**	**%**	**Count**	**%**	**Count**	**%**
Phonemic Errors	8	44	14	78	5	28
Phonetic Errors	2	11	13	72	0	0
Intonation	15	83	16	89	4	22
Grammatical Errors	10	56	14	78	3	17
Utterance Length	0	0	0	0	0	0

We also examined intercorrelations among the stimulus assessments, as shown in Table 3. Correlations significant at $p < .05$ are marked with an asterisk. The number of grammatical errors was significantly correlated with both the number of phonemic errors and the number of phonetic errors. In addition, the intonation ratings were correlated with phonemic error scores. In general, speakers who made grammatical errors also tended to make pronunciation errors. The surprising lack of correlations between phonemic and phonetic errors and phonetic errors and intonation suggests that errors in each of these categories were independent of one another.

Table 3. Intercorrelations (Pearson *r*) of stimulus characteristics [Recreated]

	Phonetic	Intonation	Grammar
Phonemic	.22	.39*	.48*
Phonetic		.23	.39*
Intonation			.28

* $p < .05$

Discussion

… Native English listeners transcribed and rated for comprehensibility and foreign accent a set of speech samples produced by 10 proficient ESL learners. Overall, they found the nonnative stimuli to be highly intelligible. In fact, more than half of the transcriptions received scores of 100%, and many others contained only minor errors. Although the utterances also tended to be highly rated in terms of perceived comprehensibility, the range of scores on the accent rating task was quite wide, with a noteworthy proportion in the "heavily accented" range.

There are a number of reasons to suppose that the three types of scores under consideration here correspond to related but partially independent dimensions. Similar to Varonis & Gass (1982), who observed strong correlations between judgments of comprehensibility and binary good/bad pronunciation judgments, we observed strong correlations between comprehensibility and intelligibility in extemporaneous speech. However, we found a number of important differences as well. First, the distributions of perceived comprehensibility and accent scores were noticeably different; the listeners tended to assign harsher scores when rating accent. Second, the strength of the correlation among any of the three possible pairings of dimensions tended to be in the moderate range for most listeners.…. Third, far fewer listeners showed a significant correlation between intelligibility and accent than between intelligibility and perceived comprehensibility. The accent scores were a much poorer reflection of the listeners' actual comprehension of an utterance than were the perceived comprehensibility scores. Our new Figure A underscores the closer connection between comprehensibility and intelligibility in that the best comprehensibility ratings (1 or 2) tended to be associated with 100% intelligibility. No such tendency emerged for the accentedness ratings.

We found a fourth important difference when we examined a subset of the data. The listeners sometimes rated utterances as moderately or heavily accented even when able to transcribe them perfectly. This finding demonstrates empirically that the presence of a strong foreign accent does not necessarily result in reduced intelligibility or comprehensibility.

We stress that our observation should not be misinterpreted to mean that accented speech is *always* fully intelligible and comprehensible. The fact remains some aspects of an accent can negatively affect intelligibility. Our finding here points to the importance of identifying those aspects of an accent or combinations that have a deleterious effect.

The intelligibility scores were the most direct test of what the listeners actually understood, because they indicated which words in each utterance the listeners had correctly identified.… The lack of complete congruence between intelligibility and perceived comprehensibility was probably due to factors that the listeners took into account when making comprehensibility judgments but that did not necessarily determine whether an utterance was fully understood. For instance, two foreign-accented utterances may both be fully understood (and therefore be perfectly intelligible), but one may require more processing time than another. Munro and Derwing (1995b), for instance, found that L2 utterances generally took longer to process than L1 speech.… The need to allocate extra processing resources to an utterance might cause a listener to assign a lower comprehensibility score. Our conception of the relationship can best be summarized as follows: two utterances can both turn out to be fully intelligible, yet one may require the listener to follow a more demanding path to arrive at comprehension or understanding.

We assume that listeners judged accentedness as the extent to which the pronunciation of each utterance deviated from some notion of what a native-like version would be. The foreign accent scores did not predict intelligibility very well. Perhaps, when judging accentedness, listeners were primarily influenced by variables that caused the speech samples to sound deviant but that ultimately had little impact on whether the message was understood.…. This was our working hypothesis when we undertook our 2006 study of functional load, in which we found evidence that low functional load pronunciation errors detracted less from comprehensibility than did high functional load errors (see also, Kang & Moran, 2014).

… Unlike many other studies, ours used extemporaneous utterances rather than excerpts from reading passages or sentence stimuli. As a result, we examined accent and intelligibility under circumstances that better reflect naturally occurring speech. Thus, this study addresses the relationship between accent and intelligibility more directly. In later studies however, we assessed the same three dimensions in controlled utterances using identical techniques (Derwing & Munro, 1997; Derwing, Munro & Wiebe 1998, Derwing et al., 2014). Ultimately, though, extemporaneous and spontaneous speech, both monologic and interactive in nature, are more representative of actual communication than read speech. Thomson and Derwing (2015), in a narrative review of 75 pronunciation intervention studies, noted that 73% employed read speech samples, while 12% used elicited imitation. In the interests of ecological validity, it would be useful to further extend comparisons of a variety of speaking contexts including interactive speech (Crowther, 2020; Derwing, Rossiter, Munro, & Thomson, 2004; Trofimovich et al., this volume; Zielinksi & Pryor, this volume) to measure change in pronunciation.

Implications for second language teaching and research

These findings have important implications for pronunciation assessment and instruction for adult second language learners. As far as we know, these are the first experimental data demonstrating what pronunciation experts have long believed: Although strength of foreign accent is indeed correlated with comprehensibility and intelligibility, a strong foreign accent does not necessarily cause L2 speech to be low in comprehensibility or intelligibility.… The nature of the scale to be used in assessment should be determined according to the goals of the instructor and the learner. If comprehensibility and intelligibility are accepted as the most important goals of instruction in pronunciation, then the degree to which a particular speaker's speech is accented should be of minor concern, and instruction should not focus on global accent reduction, but only on those aspects of the learner's speech that appear to interfere with listeners' understanding.

This raises two problems for those who teach pronunciation to second language learners. First, at present little empirical evidence indicates which particular aspects of foreign-accented speech are most detrimental to comprehensibility and intelligibility. Since Munro and Derwing (1995a), research has pointed to some elements of accent that can interfere with listener understanding. Hahn (2004) demonstrated that both monotone and inappropriately placed primary stress detract from intelligibility. To our knowledge, Munro and Derwing (2006) were the first to empirically test the hypothesis (based on Catford, 1987) that high functional load (FL) segmental errors would result in greater difficulty for comprehensibility than low functional load errors. That exploratory study suggested that indeed, high FL errors had a cumulative negative effect on listeners' comprehensibility ratings, while low FL errors did not. Speech rate also had a limited but significant influence on comprehensibility. Munro and Derwing (2001) determined that speech that is either too slow or too fast can cause problems for listeners. Comprehensibility can also be compromised if L2 speakers do not follow local pragmatics conventions. Many speech acts are highly predictable, so if an L2 speaker uses unexpected patterns, additional pressure is put on listeners (Derwing, Waugh & Munro, 2021). All of these studies point to possible interventions for L2 learners, but more research is needed to identify appropriate interventions and evaluate their outcomes.

Second, there are individual differences in the perception of nonnative speech. Although our listeners tended to agree among themselves in their judgments, there were also important individual differences in the relationships among accentedness and comprehensibility ratings and intelligibility scores. It follows that opinions of a particular speaker's most serious pronunciation problems may vary from listener to listener. There are a number of possible explanations for the variability in this study. First, individual listeners may have interpreted the instructions differently. Some, for instance, may have focused more on the syntactic properties of the stimuli than others. Second, familiarity with accented speech may have influenced some listeners' results (cf. Gass & Varonis, 1984…). Only one listener reported having any regular contact with Mandarin speakers (that person's orthographic transcription score was well below the mean); however, of the six people who reported having fairly frequent contact with more than one other accent, five had orthographic transcription scores above the mean. As pointed out earlier with respect to NSs, individuals may vary in terms of rate of speech, speech clarity, voice quality, word choice, control of pragmatic conventions, and so forth. All of these variables affect the comprehensibility of NNSs' speech as well. Finally, irrespective of differences in experience with L2 speech, there are probably individual differences in the ability to comprehend it.

Clearly, we need further studies of those aspects of L2 pronunciation that have the greatest impact on intelligibility. This study dealt only with one variety of accent (Mandarin), and the samples were elicited from individuals who were all proficient in English. Studies that include a variety of accents produced by speakers with differing levels of proficiency (e.g., Derwing & Munro, 1997), and that give attention to differences among raters (e.g., Munro, Derwing & Holtby, 2012) should help to elucidate the relative contributions to intelligibility of specific elements (subsegmental, segmental, prosodic) of pronunciation. For instance, our study shows that intonation figures importantly in listener judgments of comprehension and accent, at least for Mandarin speakers of English. In addition, a recent study by Anderson-Hsieh et al. (1992) has provided promising empirical evidence in favor of prosody as a factor in the intelligibility of L2 speech, but this work must still be regarded as preliminary, given that a clear distinction between accent and intelligibility was not made. Theoretical analysis by Catford (1987) on functional load in English may provide a direction for future studies at the segmental level. (See comments above regarding functional load.) We ourselves plan to explore this issue in more detail by examining how accent and intelligibility are related to other variables, such as processing time (see Munro & Derwing, 1995b) and subjective listener reactions to nonnative pronunciation. In Derwing and Munro (2009) and Derwing (2016), we explored engineers' reactions to nonnative speech in a preference task; both comprehensibility and fluency factored into their choices.

Retrospective interpretations

When we reflect on what our 1995 study achieved, several points come to mind:

1. The partial independence of the three dimensions and the high reliability of listener ratings have been demonstrated repeatedly since 1995 in studies of our own and of numerous other researchers; our three-way model thus provides a good framework for describing L2 pronunciation. From a practical standpoint, we see it as essential that pronunciation teachers learn about this research evidence as part of their training. At the same time they should be made aware that instructional studies have demonstrated that comprehensibility can improve without a change in accentedness (e.g., Derwing, Munro & Wiebe, 1998; Gordon, 2021).
2. The study yielded a framework that has been applied in longitudinal research and in intervention studies to provide us and others with a good tool for measurement of L2 speech and pronunciation learning (Derwing & Munro, 2013; Derwing, Munro & Wiebe, 1998; Gordon, 2021; Zielinski & Pryor, this volume; Zhang & Yuan, 2020). We realized that one-shot studies such as this one cannot probe change over time, so we later conducted a ten-year longitudinal study of naturalistic pronunciation development (e.g., Derwing & Munro, 2013).
3. A reviewer of the original manuscript insisted that "comprehensibility" was not an acceptable abbreviation for "perceived comprehensibility." Accord-

ingly, we changed the wording in several places. We regret not having stood up to this reviewer, because it is now very clear to us that there can be no type of comprehensibility other than "perceived comprehensibility." Comprehensibility must be operationalized in terms of listener responses; the notion is meaningless otherwise. The same applies to accentedness. On the related issue of standardization of terminology, we believe that usage has become somewhat more consistent, despite occasional confusions. These problems can probably be alleviated if researchers are careful to specify the terminological definitions they are assuming.

4. We caution researchers that scalar ratings on the accentedness and comprehensibility scales are not norm-referenced and therefore can be interpreted only in a relative sense. While Munro (2018) reported that listeners treated a 9-point scale as an equal-interval dimension, Derwing and Munro (2009) appear to be the only researchers to have found that a single-point difference in ratings from one study can be reliably perceived by a different group of listeners from another. More research is necessary in this area (see Nagle & Huensch, this volume).
5. Despite some advances in our understanding of the causes of intelligibility breakdowns in L2 speech, much more empirical work remains to be done, particularly with respect to identifying instructional approaches that bring about long-term improvement in intelligibility.
6. Approaches to data analysis have evolved since 1995, in part because of the availability of improved software tools and hardware capabilities. Nagle and Huensch (this volume) inspired us to use mixed-effects modelling (albeit much simpler than theirs) to examine the relationship between intelligibility and the two other dimensions in the 1995 data. Accordingly, we followed their approach of recoding our intelligibility scores as binary values, except that we classified utterances as either *100% intelligible* or *less than 100% intelligible* (as shown earlier in Figure A). This yielded a relatively even distribution of the two possibilities. We then fit a series of mixed effects binary logistic models with the *glmer* function of *lmer4* (Bates et al., 2015) in *R*. Accentedness and comprehensibility were fixed predictors, with listeners as a random effect. (Because of incomplete data we could not include a random speaker effect. Moreover, given the small sample size, we consider this analysis purely exploratory.) Models based on only comprehensibility and only accentedness both outperformed a null model, χ^2 (1) = 108.01 and 26.8, respectively, ps < .001. However, removing accentedness from a model based on both predictors resulted in no significant change χ^2 (1) = .01, p = .91. Following Nagle and Huensch we computed odds ratios for comprehensibility-only and accentedness-only models by exponentiating from log odds. The results, sum-

marized in Table A, suggest that, while both comprehensibility and accentedness ratings could predict to some degree whether or not a particular utterance was intelligible, comprehensibility (odds ratio = 1.754) outperformed accentedness (1.248) as a predictor.

Table A. Odds ratios for generalized mixed effects binary logistic models fit to intelligibility scores (including listeners as a random effect)

			95% CI		
Model	**Effect**	**Estimate**	**Lower**	**Upper**	***p***
Comprehensibility only	Intercept	0.154	0.096	0.245	<.001
	Comprehensibility	1.754	1.542	1.996	<.001
Accentedness only	Intercept	0.288	0.173	0.479	<.001
	Accentedness	1.248	1.142	1.363	<.001

7. Our hope upon embarking on this study and pursuing our subsequent work was to contribute to an understanding of the nature of L2 speech by identifying ways to enhance L2 learners' communication skills. It is a well-established fact that an accent can elicit discriminatory behaviour. For that reason, it is a cause of some distress to us that some researchers continue to conduct studies to rediscover the phenomenon of accent discrimination. Our advice to them is to *Get over it* and *Get on with it*! A focus on identifying accent discrimination in the absence of suggestions for ways to overcome it has little practical importance. We know, empirically, that listeners prefer speech that is easy to understand, regardless of the degree of accentedness (Derwing & Munro 2009). We also know that comprehensibility can be enhanced without changing accentedness (Derwing, Munro & Wiebe, 1998; Gordon, 2021). Furthermore, comprehensibility and intelligibility are far more important to L2 speakers' overall welfare than accent. For instance, we would like to see more research focusing on what listeners can contribute to the success of interactions (Derwing, 2016; Derwing, Rossiter & Munro, 2002; Kang, Rubin & Lindemann, 2015; Lindemann, Campbell, Litzenberg & Subtirelu, 2016). The studies that have attempted to assist native speakers to engage with accented speakers are limited in range; countless other approaches are possible but need exploration. In the next twenty-five years, we hope to see a stronger focus on addressing social problems associated with L2 speech with practical solutions.

Acknowledgements

In 1995, we acknowledged financial support from a grant from the University of Alberta to the second author and a SSHRC postdoctoral fellowship to the first author. We also expressed appreciation to J. Flege, B. Derwing, A. Schmidt, H. Southwood, N. Takagi and three anonymous reviewers for their comments on earlier drafts. We are grateful to the Editors of *Language Learning* for allowing us to reprint substantial portions this article, which first appeared in Volume 45(1) and was subsequently reprinted in a special issue of *Language Learning* in 1999; Alistair Cumming, the Editor of *Language Learning* at the time, was very supportive of us both. Thanks are due to Bruce Derwing, John Levis and Charlie Nagle for their comments on this new version. Over the years we have benefitted from the assistance of several students and colleagues too numerous to mention, but we especially acknowledge Marian Rossiter and Ron Thomson. Finally, we are indebted to the people who were the impetus for our curiosity about accent, comprehensibility, and intelligibility: the Vietnamese students in the ESL classes we taught when we were graduate students. Their tenacity and resilience were inspiring.

References

Abercrombie, D. (1949). Teaching pronunciation. *English Language Teaching*, 3, 113–122. https://doi.org/10.1093/elt/III.5.113

Anderson-Hsieh, J., Johnson, R., & Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning*, 42, 529–555. https://doi.org/10.1111/j.1467-1770.1992.tb01043.x

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Brennan, E. M., & Brennan, J. S. (1981). Accent scaling and language attitudes: Reactions to Mexican American English speech. *Language & Speech*, 24, 207–221. https://doi.org/10.1177/002383098102400301

Catford, J. C. (1987). Phonetics and the teaching of pronunciation: A systemic description of English phonology. In J. Morley (Ed.), *Current perspectives on pronunciation: Practices anchored in theory* (pp. 83–100). Washington, DC: TESOL.

Crowther, D. (2020). Rating L2 speaker comprehensibility on monologic vs. interactive tasks: What is the effect of speaking task type? *Journal of Second Language Pronunciation*, 6(1), 96–121. https://doi.org/10.1075/jslp.19019.cro

Derwing, T. M. (2016). The 3 P's of ESL in the workplace: Proficiency, pronunciation and pragmatics. In H. McGarrell & D. Wood (Eds.) *Contact, Refereed Proceedings of the TESL Ontario Research Symposium*, 42 (2), 10–20.

Derwing, T. M., & Munro, M. J. (1997). Accent, comprehensibility and intelligibility: Evidence from four L1s. *Studies in Second Language Acquisition*, 19, 1–16. https://doi.org/10.1017/S0272263197001010

Derwing, T. M. & Munro, M. J. (2009). Comprehensibility as a factor in listener interaction preferences: Implications for the workplace. *Canadian Modern Language Review*, 66 (2) 181–202. https://doi.org/10.3138/cmlr.66.2.181

Derwing, T.M. & Munro, M.J. (2013). The development of L2 oral language skills in two L1 groups: A seven-year study. *Language Learning*, 63, 163–185. https://doi.org/10.1111/lang.12000

Derwing, T.M., Munro, M.J., Foote, J.A., Waugh, E. & Fleming, J. (2014). Opening the window on comprehensible pronunciation after 19 years: A workplace training study. *Language Learning*, 64, 526–548. https://doi.org/10.1111/lang.12053

Derwing, T.M., Munro, M.J. & Thomson, R.I. (2008). A longitudinal study of ESL learners' fluency and comprehensibility development. *Applied Linguistics*, 29, 359–380. https://doi.org/10.1093/applin/amm041

Derwing, T.M., Munro, M.J. & Wiebe, G.E. (1998). Evidence in favor of a broad framework for pronunciation instruction. *Language Learning*, 48, 393–410. https://doi.org/10.1111/0023-8333.00047

Derwing, T.M., Rossiter, M.J., & Munro, M.J. (2002). Teaching native speakers to listen to foreign-accented speech. *Journal of Multilingualism and Multicultural Development*, 23, 245–259. https://doi.org/10.1080/01434630208666468

Derwing, T.M., Rossiter, M.J., Munro, M.J. & Thomson, R.I. (2004). L2 fluency: Judgments on different tasks. *Language Learning*, 54, 655–679. https://doi.org/10.1111/j.1467-9922.2004.00282.x

Derwing, T.M., Waugh, E., Munro, M.J. (2021). Pragmatically speaking: Preparing adult ESL students for the workplace. *Applied Pragmatics*, 3, 107–135.

De Weers, N. (2020). A critical (re)assessment of the effect of speaker ethnicity on speech processing and evaluation. (Unpublished doctoral dissertation). Department of Linguistics, Simon Fraser University, Burnaby, BC, Canada.

Fayer, J.M., & Krasinski, E. (1987). Native and nonnative judgments of intelligibility and irritation. *Language Learning*, 37, 313–326. https://doi.org/10.1111/j.1467-1770.1987.tb00573.x

Gass, S.M., & Varonis, E.M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning*, 34(1), 65–89. https://doi.org/10.1111/j.1467-1770.1984.tb00996.x

Gordon, J. (2021). Pronunciation and task-based instruction: Effects of a classroom intervention. *RELC Journal Special Issue*, 52(1), 94–109. https://doi.org/10.1177/0033688220986919

Hahn, L.D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly*, 38(2), 201–223. https://doi.org/10.2307/3588378

Kang, O. & Moran, M. (2014). Functional loads of pronunciation features in non-native speakers' oral assessment. *TESOL Quarterly*, 48, 176–187. https://doi.org/10.1002/tesq.152

Kang, O., Rubin, D., & Lindemann, S. (2015). Mitigating U.S. undergraduates' attitudes toward international teaching assistants. *TESOL Quarterly*, 49, 681–706. https://doi.org/10.1002/tesq.192

Lindemann, S., Campbell, M.-A., Litzenberg, J., & Subtirelu, N.C. (2016). Explicit and implicit training methods for improving native English speakers' comprehension of nonnative speech. *Journal of Second Language Pronunciation*, 2, 93–107. https://doi.org/10.1075/jslp.2.1.04lin

Lindemann, S., & Subtirelu, N. (2013). Reliably biased: The role of listener expectation in the perception of second language speech. *Language Learning*, 63(3), 567–594. https://doi.org/10.1111/lang.12014

Munro, M.J. (1995). Nonsegmental factors in foreign accent: Ratings of filtered speech. *Studies in Second Language Acquisition*, 17 (1), 17–34. https://doi.org/10.1017/S0272263100013735

Munro, M.J. (2018). Dimensions of pronunciation. In O. Kang, R. Thomson, & J. Murphy. *The Routledge Handbook of Contemporary English Pronunciation*. pp. 413–431. New York: Routledge.

Munro, M.J. & Derwing, T.M. (1995a). Foreign accent, comprehensibility and intelligibility in the speech of second language learners. *Language Learning*, 45, 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Munro, M.J., & Derwing, T.M. (1995b). Processing time for native and foreign accented speech . *Language & Speech*, 38, 289–306. https://doi.org/10.1177/002383099503800305

Munro, M.J., & Derwing, T.M. (1994). Evaluations of foreign accent in extemporaneous and read material. *Language Testing*, 11, 253–266. https://doi.org/10.1177/026553229401100302

Munro, M.J. & Derwing, T.M. (2001). Modelling perceptions of the comprehensibility and accentedness of L2 speech: The role of speaking rate. *Studies in Second Language Acquisition*, 23, 451–468. https://doi.org/10.1017/S0272263101004016

Munro, M.J. & Derwing, T.M. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System*, 34, (520–531). https://doi.org/10.1016/j.system.2006.09.004

Munro, M.J., Derwing, T.M., & Holtby, A.K. (2012). Evaluating individual variability in foreign accent comprehension. In J. Levis & K. LeVelle (Eds.). Proceedings of the 3rd Pronunciation in Second Language Learning and Teaching Conference, Sept. 2011, (pp. 233–239). Ames, IA: Iowa State University.

O'Brien, M.G. (2016). Methodological choices in rating speech samples. *Studies in Second Language Acquisition*, 38(3), 587–605. https://doi.org/10.1017/S0272263115000418

Rondeau, G. (1972). *Le français international, Livre* 2. Montréal: Centre Educatif et Culturel.

Schairer, K. (1992). Native speaker reaction to non-native speech. *Modern Language Journal*, 76, 309–319. https://doi.org/10.1111/j.1540-4781.1992.tb07001.x

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86, 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Thomson, R.I. & Derwing, T.M. (2015). The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics*, 36, 326–344. https://doi.org/10.1093/applin/amu076

Van Els, T. & De Bot, K. (1987). The role of intonation in foreign accent. *Modern Language Journal*, 71, 147–155. https://doi.org/10.2307/327199

Varonis, E.M., & Gass, S.M. (1982). The comprehensibility of nonnative speech. *Studies in Second Language Acquisition*, 4, 114–136. https://doi.org/10.1017/S027226310000437X

Zhang, R. & Yuan, Z. (2020). Examining the effects of explicit pronunciation instruction on the development of L2 pronunciation. *Studies in Second Language Acquisition*, https://doi.org/10.1017/S0272263120000121

Zielinski, B. (2008). The listener: No longer the silent partner in reduced intelligibility. *System*, 36(1), 69–84. https://doi.org/10.1016/j.system.2007.11.004

Revisiting the Intelligibility and Nativeness Principles[*]

John Levis
Iowa State University

Levis (2005) named two conflicting approaches to pronunciation teaching, the Nativeness Principle and the Intelligibility Principle. This paper revisits those two principles to argue for the superiority of the Intelligibility Principle in regard to where pronunciation fits within the wider field of language teaching, in how it effectively addresses teaching goals, in how it best addresses all contexts of L2 pronunciation learning, and in how it recognizes the reality of social consequences of pronunciation differences. In contrast, the Nativeness Principle, despite its long pedigree and many defenders, falls short by advocating native pronunciation for L2 learners, which is both unlikely to be achieved and unnecessary for effective communication in the L2.

Keywords: Intelligibility principle, Nativeness principle, pronunciation teaching, social factors, World Englishes

1. Introduction

In 2005, I was the guest editor for a special issue of *TESOL Quarterly* titled "Reconceptualizing Pronunciation in TESOL: Intelligibility, Identity, and World Englishes." To help frame the special issue, I described a conflict that had long been simmering within the field of L2 pronunciation (Levis, 2005). I described the conflict in terms of two approaches to pronunciation teaching, which I named the Nativeness Principle and the Intelligibility Principle. This article has been cited over 1000 times, and the Nativeness Principle and the Intelligibility Principle have become part of the way we talk about approaches to the teaching and learning of pronunciation.

[*] An earlier version of this article was published as part of a special issue of the *Journal of Second Language Pronunciation* 6(3). https://doi.org/10.1075/jslp.20050.lev

https://doi.org/10.1075/bct.121.03lev

Because our beliefs about pronunciation (reflected in the two principles) have consequences for how we teach and learn pronunciation, it was my argument that the Intelligibility Principle better matched the reality of learning L2 pronunciation. Many, however, still treat the Nativeness Principle as a valid alternative view of teaching pronunciation, so it is worth revisiting the two principles to update our understanding. Even though I argued about the state of English pronunciation teaching, it is now clear that issues relevant to English are equally relevant to most other languages as well. As a result, this paper is about intelligibility and nativeness in language teaching, not just in relation to English. In revisiting the 2005 article, I will argue that the Intelligibility Principle is consistent with what we know about L2 pronunciation learning, while the Nativeness Principle is deeply faulty in its approach to L2 pronunciation. It is faulty in how it relates L2 pronunciation to L2 language learning in general, in what it implies for teaching and learning goals, in its inability to address all contexts of pronunciation learning, and in how it addresses social aspects of pronunciation.

2. Terminology in Levis (2005) and Munro and Derwing (1995)

This volume highlights the centrality of Munro and Derwing (1995) to pronunciation research and teaching, and especially the influence of their constructs of intelligibility, comprehensibility and accentedness. (In this volume, the authors reconsider their earlier paper and its findings, providing new analyses that strengthen the centrality of the original research to today's field.) In revisiting the Intelligibility and Nativeness Principles, it is important to connect my two principles to Munro and Derwing's terms (see Table 1). In my 2005 article, I used the word "intelligibility" quite generally, in the sense used by Merriam Webster, "capable of being understood or comprehended." My use of intelligibility thus implies both actual understanding (*intelligibility* in Munro & Derwing, 1995) and the ease with which understanding occurs (*comprehensibility* in Munro & Derwing, 1995). In contrast, my Nativeness Principle addressed only the issue of *accentedness* as used by Munro and Derwing (Table 1). The Nativeness Principle seems to assume that speakers will be both intelligible and comprehensible if they match a native model, but this is only implicit. Explicitly, intelligibility and comprehensibility are extraneous to a view that prioritizes nativeness.

Table 1. Relation of Terms Used in Munro and Derwing (1995) and Levis (2005)

		Principles in Levis (2005)	
		Nativeness Principle	**Intelligibility Principle**
Munro and Derwing (1995) terms	Accentedness	Central to Nativeness	Largely irrelevant
	Intelligibility	Not explicitly discussed	Actual Understanding
	Comprehensibility	Not explicitly discussed	Ease of Understanding

3. Nativeness, Intelligibility and Pronunciation Teaching

Ideologies of nativeness and near-nativeness are deeply entrenched within L2 pronunciation, partly because of the influence of Chomsky's (1965) concept of competence, or what hypothetical ideal (native) speaker/listeners know, that is, their knowledge about the language. As a result, nativeness has frequently been used to describe how those who are not monolinguals (e.g., bilinguals and L2 learners) differ from monolinguals, with native monolinguals usually setting the standard. The second part of Chomsky's formulation, performance, involved what ideal speaker/listeners actually do when they use language in real time. Although of little interest to Chomsky, L2 teachers and learners live in a world of performance. Research has shown that L2 users and bilinguals may have native-like performance in various aspects of the L2 but that they typically do not have the same language knowledge representations (i.e., competence) as monolingual native speakers (e.g., Coppieters, 1987; Sorace, 1993). These findings show the vast differences between Chomsky's ideal speaker/listener with a monolingual grammar and the reality for L2 learners (e.g., see Sorace, 2003 for a discussion of near-nativeness), especially in regard to pronunciation (e.g., Sakai, 2018), in which performance is central.

Among language learners, many think it possible to sound like a native speaker. Indeed, that is the desire of many, especially among immigrants in inner circle countries. However, in language teaching, privileging nativeness or near-nativeness has been widely criticized, and nativeness has very little currency as an ultimate goal for L2 learning (Agudo, 2017). Indeed, there is consensus among professional language teaching organizations that there is no justification to privileging native speaker identity or demanding near-native performance in any context of language teaching (e.g., https://www.tesol.org/docs/pdf/5889.pdf). That we are still talking about the Nativeness Principle in regard to pronunciation teaching shows that pronunciation teaching has often been out of touch with the wider concerns of L2 teaching and learning.

A possible reason that the Nativeness Principle remains alive and well in pronunciation teaching is that pronunciation teaching and learning have been neglected since the advent of the communicative era (Levis & Sonsaat, 2017). As a result, pronunciation has developed separately from other aspects of language teaching, and the Nativeness Principle continues to be an attractive goal for many teachers and learners. Unfortunately, the Nativeness Principle actually assumes things that are largely unattainable (e.g., that adult learners can become nativelike in pronunciation) and unnecessary (e.g., that nativeness is necessary for communicative success). The evidence for why nativeness is usually unattainable and unnecessary is addressed in Section 4.

4. Nativeness, Intelligibility and their Implications for Pronunciation Teaching

In 2005, I talked about the Intelligibility Principle and Nativeness Principle as being "contradictory" (p. 370). By this, I meant that the two principles were rooted in fundamentally different approaches to language teaching even though the practices associated with the two principles often overlapped and looked similar. For example, even though both approaches agree on the importance of pronunciation for language teaching, and both are likely to prioritize certain features and use similar techniques, they differ in their evaluation of student success, in decisions about who is a qualified teacher, and in how they talk about success. Like the famous poem by Robert Frost, the principles are two roads that diverge, and following one road precludes traveling on the other (https://www.poetryfoundation.org/poems/44272/the-road-not-taken).

My argument was, and is, that the Intelligibility Principle is a superior way to think about pronunciation teaching and learning. It is more in line with what we know about ultimate attainment in L2 pronunciation, it recognizes that diversity in accentedness is only very indirectly related to impaired communication and that speakers who are perceived as strongly accented can also be highly intelligible (Munro & Derwing, 1995; Derwing & Munro, 2015), it honors the abilities of all qualified language teachers and recognizes the great strengths that nonnative teachers bring to the teaching of pronunciation, and it recognizes that not all pronunciation features are equally important. Far from promoting a "limited degree of phonological competence" (Pennington & Rogerson-Revell, 2019, p. 132), the Intelligibility Principle better reflects the reality of accent diversity in English (indeed, in any world language and L2 context). The Nativeness Principle, on the other hand, has always been based on a myth that there are ideal and deficient ways to pronounce a language, and that deficient ways to pronounce should not be tolerated. As a result of these divergent beliefs, the Nativeness and Intelligibility Principles also diverge in how they address pedagogical issues, in who they consider to be an ideal teacher, and in how they accommodate accent diversity.

With reference to pedagogically-oriented issues, the Nativeness Principle is deeply problematic because it assumes that all aspects of pronunciation are, de facto, equally important, and that no matter where a learner starts, there is only one allowable destination: sounding like a native speaker. Any unmastered pronunciation feature demonstrates that the learner has failed. In contrast, the Intelligibility Principle asserts that communicative success, not nativeness, is the goal, and that not all pronunciation features are equally important for being understood. For example, L2 consonant or vowel contrasts are sometimes important based on the functional load of the contrasts (Brown, 1988). Functional load is a measure of the likelihood that two sounds will be confused by listeners. There is compelling evidence that errors in higher functional load segmental features are associated with greater loss of comprehensibility, which in Section 2 above is part of the Intelligibility Principle (Munro & Derwing, 2006; Suzukida & Saito, 2019). In addition, suprasegmental features such as prominence placement can lead to worse comprehension for listeners (Hahn, 2004) while some stress and intonational features do not appear to affect understanding in the same way (Cutler, 1986; Levis, 1999).

A second assumption of the Nativeness Principle is that only teachers who are native or native-like can be trusted to teach pronunciation. A focus on nativeness leaves many well-qualified nonnative teachers uncertain of whether they should teach pronunciation or trust their own skills. If they want to teach pronunciation, they may be seen as deficient models of L2 speech by their students, their colleagues or even themselves. Believing that nativeness is a realistic standard for L2 learning can also foster discriminatory practices because nonnative teachers may be considered deficient native speakers (Mahboob & Golden, 2013; Medgyes, 1992). This is especially true for pronunciation. Well-qualified L2 speakers may be passed over as teachers of oral skills (including pronunciation), and native speakers may be prioritized for teaching opportunities simply because they are native (Buckingham, 2015; Moussu & Llurda, 2008). On the other hand, the Intelligibility Principle recognizes that being a native speaker is neither a necessary nor sufficient qualification to teach L2 pronunciation. Rather than elevating nativeness as the primary qualification, the Intelligibility Principle recognizes that L2 pronunciation is best taught by qualified language teachers, and that nativeness is not a required or even a preferred qualification when it comes to student learning (Levis et al., 2016). It also recognizes that nonnative teachers bring unusually strong skills to the teaching of pronunciation because of their own history of learning the pronunciation of the L2.

A third implication of the Nativeness Principle is that only certain native accents (such as General American or Standard Southern British when considering English) are truly acceptable. In other words, many native speakers are likely to find themselves on the outside of a club that privileges certain ways of speaking and ignores or denigrates others. In contrast, the Intelligibility Principle proposes that a wide variety of accents are acceptable as teaching models and that speakers need not converge only toward prestige accents. Teachers and learners can use or develop their own accents, adjusting them as needed in different contexts to achieve intelligibility. Any language in which pronunciation is taught is enriched by its multiple accents, and a wider familiarity with these accents may also promote the ability to interact and understand other speakers (Major et al., 2002; Ockey & French, 2016). In languages like Spanish, Arabic, French and Hindi, which have many different regional and social accents, there is tremendous mutual intelligibility despite the diversity of accents. Even though there may be powerful social biases toward certain varieties, L2 learners should not be made party to L1 language prejudices if they are intelligible. The ability to understand several accents occurs because of the flexibility of human listeners (Scharenborg, 2007) and because humans are very good at adapting to unfamiliar native (Adank, Evans, Stuart-Smith & Scott, 2009) and nonnative speech patterns (Baese-Berk, Bradlow & Wright, 2013). When pronunciation is intelligible (in the broad sense, that is, including both intelligibility and comprehensibility), then the Intelligibility Principle says that it does not need to be taught.

5. How are the Nativeness and Intelligibility principles related?

The relationship between the Nativeness and Intelligibility principles can be visualized in terms of how they overlap and what they say about the relative importance of pronunciation in communication. If the two principles are seen only as two ways to talk about pronunciation, intelligibility will inevitably be seen as an abridged form of Nativeness (Figure 1) in which not all pronunciation features included in nativeness are included in intelligibility, though all aspects of intelligibility are part of nativeness. This perhaps corresponds to a belief that intelligibility reflects reduced standards.

One reason why this view of intelligibility is faulty is because it assumes that speech intelligibility is simply a matter of pronunciation. Research demonstrates that intelligibility includes more than pronunciation (e.g., Jenkins, 2000, in which two-thirds of interactions with lost intelligibility were connected to pronunciation

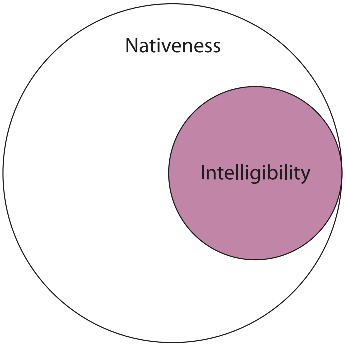


Figure 1. Intelligibility as reduced pronunciation requirements

while one-third were related to vocabulary and grammar). Figure 1 is also unsatisfactory because of what it implies about teaching pronunciation. It implies that the Nativeness Principle upholds higher standards of performance and knowledge while the Intelligibility Principle chooses to ignore much of what is known about a language's pronunciation. However, those who advocate intelligibility do so not because they advocate reduced standards but rather because communicative success does not require most of what can be taught about pronunciation. Language learners are not required to become expert phoneticians to communicate.

The relationship between the Nativeness and Intelligibility principles can also be visualized as one of some overlap in which Intelligibility is partially concerned with issues of pronunciation (Figure 2). In this image, intelligibility overlaps with nativeness in pronunciation, but intelligibility also involves other aspects of language (implied by the non-overlapping area) that impact communicative effectiveness such as lexical choice, grammatical accuracy, and sociolinguistic appropriateness (e.g., Jenkins, 2000). In most respects, this is a workable if incomplete image of the relationship between the two principles. It demonstrates that pronunciation is essential to intelligibility; it also shows that for pronunciation teaching and learning, our goals are to identify those areas in which the two circles overlap, and emphasize those features needed by learners. The overlapping of the circles suggest a complementary relationship between intelligibility and nativeness, with different linguistic features corresponding to each (Saito, Trofimovich & Isaacs, 2016, 2017)

Finally, the two principles can be seen in another light which prioritizes intelligibility as an overall approach to oral language (Figure 3). In this view, intelligibility is the ultimate goal in oral communication (Levis, 2018), and it affects both listening and speaking in every communicative context. The uncolored portions of the circles include aspects of nativeness that do not typically impact intelligibility. In addition, while pronunciation can be crucial to whether speakers and listeners are mutually intelligible, it is not the only factor in intelligibility. Because

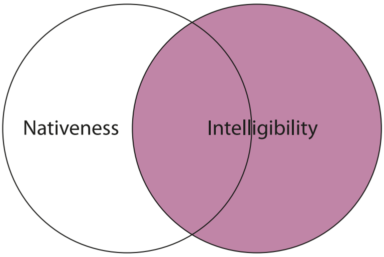


Figure 2. Intelligibility as more than pronunciation

pronunciation is an unavoidable aspect of oral communication, it is important for L2 learning insofar as it influences intelligibility. The portions of Figure 3 that overlap in multiple ways include grammatical or lexical features that are realized in their pronunciation (e.g., the different pronunciations of the -ed morpheme in English). The section titled "Other Linguistic and Non-Linguistic Aspects of Language" does not overlap with pronunciation, grammar and lexicon only because there is almost no research on how other features of communication (e.g., pragmatic appropriateness, non-verbal backchanneling, gestures, visual cues) interact with the areas that we know affect intelligibility. There is likely to be overlap. In addition, we know that there are other non-language reasons that intelligibility is impaired, such as noise, inattention, and misinterpretation of contextual clues.

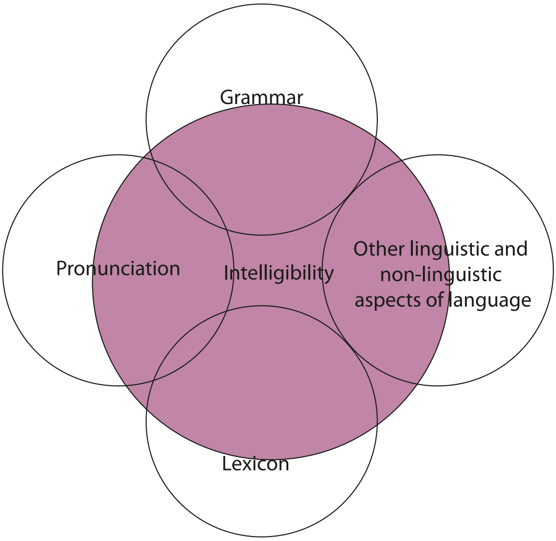


Figure 3. Intelligibility as central to oral communication

6. Research and the nativeness principle

Although nativeness may be a desired goal for specific L2 learners, the nativeness principle has very little research evidence to support it. For adult L2 learners, the age at which they began learning the L2 has a strong effect on their ultimate success. Nowhere is this effect more evident than the almost inevitable presence of a foreign accent in adult L2 learners (Flege, Munro & MacKay, 1995). Whether accents are due to factors related to age of learning (Piske et al., 2001), inadequate language experience with the L2 compared to the L1 (Bohn & Munro, 2007), the effects of identity (McCrocklin & Link, 2016), or the inability to perceive and produce L2 sounds (Kartushina & Frauenfelder, 2014), adult L2 learners only rarely become nativelike in their L2 accent.

The desire for nativeness in pronunciation often is based on beliefs that native-like speech will ensure that communication is successful (LeVelle & Levis, 2014), that learners will be more confident and respected (Derwing, 2003), that it will provide opportunities for professional advancement (Harrison, 2013), especially for language teachers (Munro, Derwing & Sato, 2006), and that it will minimize discrimination (Derwing & Munro, 2009). While these beliefs are all seem appealing, there is no evidence for the promises implied in the beliefs about developing a native accent. Likewise, the accent reduction industry, which implies similar promises for L2 learners who become more native, will not by itself get rid of discrimination (Thomson, 2014).

I have repeatedly heard researchers and teachers (including myself) say that they are in favor of aiming for intelligibility, but that if learners want to become native-like, they would encourage their attempts. This is somewhat disingenuous since we know that obtaining native-like pronunciation is highly unlikely, and that attempts to achieve this goal have two possible outcomes: Success (in extremely rare cases) and failure (in almost all cases). As a field, we should simply stop encouraging such unlikely and unnecessary goals and learn to speak of pronunciation improvement in ways that do not include myths about native-like pronunciation attainment.

Are there times that it is best to try for nativeness in pronunciation training? Yes. But the situations in which nativeness is required are few. Nativeness may be especially valued for actors who need to pass to play particular roles, such as American English speakers using a British accent (Tan, 2020) or French speakers trying to pass as English speakers in order to be considered for certain roles in English-medium films (Cerreta & Trofimovich, 2018). Nativeness may also be desired in language revitalization contexts. Bird (2020) discusses this in the context of SENĆOŦEN, a West Salish language undergoing revitalization in western Canada. Native speakers of the language are rare, and the L2 speakers both want to speak the language like the elders who still speak SENĆOŦEN but at the same time mark SENĆOŦEN as distinct from English, the dominant language. Bird discusses this in terms of the use of ejectives. Elders use weak ejectives, but the L2 learners in the community prefer strong ejectives because of their perceptual salience. Thus, even in this case, nativeness must be negotiated in relation to other factors in the social context.

7. Nativeness, Intelligibility and Contexts for Pronunciation Learning

The use of English around the world offers another example of why the Nativeness Principle is limited, whereas the Intelligibility Principle is not. Kachru's (1992) three circles of World Englishes usefully demonstrates the limitations of the Nativeness Principle by describing possible interactions between listeners and speakers (see also Levis, 2006).

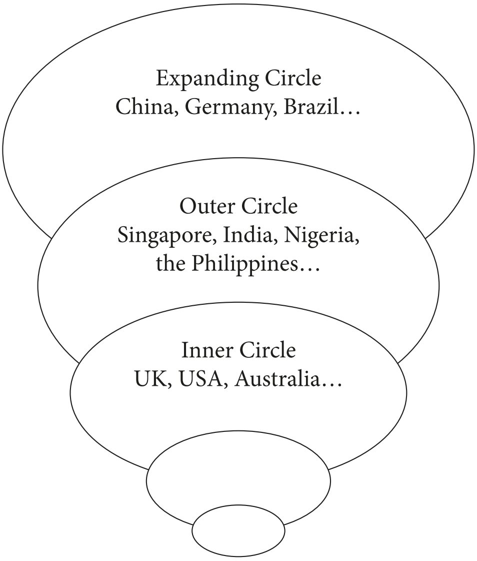


Figure 4. Three Circles of World Englishes (from Deterding, 2012)

The Inner Circle includes those who are traditionally labeled as native speakers, such as English speakers from the USA, Canada, and New Zealand. Many speakers in the Inner Circle are monolingual. The Outer Circle includes speakers from countries where English has an official role and where many people speak English regularly but as an additional language. Such countries include India, Nigeria, and Singapore. The English of speakers in these countries is not native, but rather nativized, and English is one language regularly used by multilingual speakers. Finally, Expanding Circle speakers (or nonnative speakers) come from countries where English serves as a foreign language. In Expanding Circle contexts, English has no official role and learners typically encounter it in the classroom. English is also used for tourism to most readily communicate with tourists from many countries. This means there are six options for how English speakers around the world use the language to interact (Table 2).

Table 2. Possible Intelligibility Interactions in World Englishes

	Inner Circle (IC)	**Outer Circle (OC)**	**Expanding Circle (EC)**
Inner Circle (IC)	(1) Native speakers talking to each other (e.g., Canadian and South African speakers; Southern USA and New York English speakers)	(2) Native and nativized speakers in interaction (e.g., Australian and Indian English speakers)	(3) Native and nonnative speakers in interaction (e.g., New Zealand and Japanese speakers)
Outer Circle (OC)	***************	(4) Nativized speakers talking to other Nativized speakers (e.g., Indian and Nigerian English speakers)	(5) Nativized and Nonnative speakers talking to each other (e.g., Indian and Chinese speakers
Expanding Circle (EC)	***************	***************	(6) Nonnative speakers talking to each other using English (e.g., Japanese and German speakers)

The interactions in Table 2, simplified as they are, show the limitations of the Nativeness Principle. Only (1), (2) and (3) can possibly be addressed by the Nativeness Principle, but Table 2 has nothing to say about (4)–(6), despite these types of interactions in English likely being more numerous than (1)–(3) throughout the world. In (1)–(3), the Nativeness Principle assumes that a native accent is the correct way to speak and that any loss of understanding is due to the person who is not native. As a result, the Nativeness Principle applies quite poorly to the reality of English use. At best, it can only say that everyone has to pronounce like particular native speakers, but it cannot justify such a goal beyond its implicit prejudice in favor of certain accents.

In contrast, the Intelligibility Principle is relevant for all contexts in (1)–(6). It makes no requirement that speakers with different ways of speaking have to use particular accents. It makes no claim that only certain accents will make communication possible. And finally, it recognizes that these types of interactions already take place quite successfully, and that when speakers and listeners run into trouble and certain pronunciation features are the problem, that these features should be addressed, by instruction if necessary.

There are a number of other implications from Table 2. First, intelligibility is not a matter of one person being intelligible and the other not intelligible. Instead, each speaker must be intelligible to the other. Even for native speakers talking to other native speakers (1), there is no guarantee of intelligibility. Second, both production and perception are important for an intelligibility-based approach to teaching pronunciation. Listeners must learn to understand, and speakers must speak in a way that makes them understandable. Third, preference is not automatically given to native speakers in an intelligibility-based approach. For communication to succeed, speakers must be intelligible to their listeners, whether they are other native speakers, nativized speakers, or nonnative speakers. Fourth, because there is evidence that pronunciation is important in all types of interactions in Table 2 (e.g., Jenkins, 2000; Kang, Thomson & Moran, 2018; McCullough, Clopper & Wagner, 2019; Smith & Rafiqzad, 1979), each of the contexts likely differs in how pronunciation instruction is addressed. As a result, there is no one-size-fits-all approach to teaching pronunciation.

Finally, it is important to point out that Kachru's model and the interactions between various circles relative to intelligibility and nativeness are extraordinarily simplistic in the context of expanding global mobility and digital communication. This is true not only for English but for many world languages. In fact, the interactions within each box (or between adjacent boxes) are unlikely to be limited only to those boxes. For example, this week I was in a weekly digital meeting (in English) with speakers from India, Montenegro, California, Spain, China, Thailand, and Russia. In other words, everyone now talks with everyone, via technology or through travel, so the Nativeness principle is untenable in light of this diversity of communication.

8. Intelligibility, Nativeness and Social Ramifications of Accent

The last respect in which the two principles provide different ways of understanding the importance of pronunciation is in relation to social consequences of pronunciation. The ability to distinguish accent develops early, and children under five already associate similarity or difference of accent with similarity or difference of cultural expectations (Weatherhead, White & Friedman, 2016). A wealth of previous research has shown that listeners evaluate non standard native accents more negatively than standard native accents (e.g., Dragojevic, Mastro, Giles, & Sink, 2016; Giles, Wilson & Conway, 1981; Lippi-Green, 2012). Similarly, non native accents are subject to the same kinds of negative evaluations (Gluszek & Dovidio, 2010; Harrison, 2014). Even the expectation of a non native accent may evoke socially-disadvantaged evaluations of how understandable a speaker is (Rubin, 1992).

In regard to L2 pronunciation, Pennington and Rogerson-Revell (2019) rightly recognize that "pronunciation is a social and expressive resource that can be used in conjunction with other linguistic resources to convey many different kinds of meaning" (p. 8). As a result, our beliefs about accents have social consequences for how we hear others and judge them as authentic speakers of the language. The Nativeness Principle is tightly connected to prescriptive beliefs about the social value of different accents. Choosing certain spoken varieties as pronunciation models entails a prescriptive choice by some authoritative source (even if the authority is a textbook or materials publisher). The result of the prescriptive choice ensures that the voices heard in the language classroom are limited.

The Intelligibility Principle, on the other hand, takes a descriptive view of accent variation; native and nonnative accents are in principle equal. Accent is part of the normal communicative equation, whether the interlocutors use a standard L1 accent, a nonstandard L1 accent, or an L2 accent. A descriptive view of accentedness recognizes that, by and large, native speakers adjust quickly and well to foreign-accented speakers. Clarke and Garrett (2004) found that L1-English listeners initially processed native English speech more quickly than foreign-accented speech, but that as little as a minute of exposure resulted in listeners processing foreign-accented speech more quickly. Similarly, Bradlow and Bent (2008) found that listeners were able to adjust to Chinese-accented English during the course of a presentation, and that training listeners with Chinese-accented speech helped them more successfully understand an unfamiliar Chinese-accented voice. The Intelligibility Principle is also consistent with World Englishes and English as a Lingua Franca perspectives, in which accents such as Standard Southern British and General American are simply two accents within the wider world of English accents.

Because pronunciation is always situated within a society or across social systems, those who adhere to the Intelligibility and Nativeness Principles recognize the social ramifications of accent. Both principles recognize that accent is connected to speaker identity (e.g., Gatbonton, Trofimovich & Magid, 2005), that accent may be associated with social discrimination (Lippi-Green, 2012), and that accent can overlap with issues of race and social class (Mugglestone, 1995; Subtirelu, 2015). The two principles differ, however, because of their core assumptions about language and especially about pronunciation. By providing a privileged status to particular L1 varieties, the Nativeness Principle is inherently discriminatory, even if those who adhere to it never intend to discriminate. By recognizing the validity and equivalence of different varieties, the Intelligibility Principle emphasizes successful communication across diverse accents, even if those who adhere to it sometimes treat others unequally because of the way they pronounce the language.

9. Conclusion

The Nativeness Principle and the Intelligibility Principle both continue to have defenders in the teaching and learning of L2 pronunciation. Only the Intelligibility Principle, however, accurately reflects what we know about L2 pronunciation learning and adult L2 learners. It is consistent with how the field of second language teaching understands nativeness, that is, that L2 users are not defective native speakers but multicompetent speakers in their own right (Cook, 1999). Their multicompetence includes use of grammar, lexicon, pragmatics, phonetics and phonology, as well as various types of non-linguistic, visual information such as gestures. In all respects, L2 learners do not need to be native speakers, as the Nativeness Principle assumes. The Intelligibility Principle also is consistent with realistic goals for pronunciation teaching. Whereas the Nativeness Principle asserts that L2 perfection in a particular language variety is both possible and necessary, the Intelligibility Principle recognizes that variations in accent are normal and not necessarily a barrier to communication (Derwing & Munro, 2015). The Intelligibility Principle also is relevant to all contexts of communication whereas the Nativeness Principle is not. In a world in which a massive number of interactions in varied languages take place each day without native speakers being involved, only the Intelligibility Principle recognizes the validity of contexts without native speakers. Finally, the Intelligibility Principle treats social variation in accent not as a problem to overcome but as variation to embrace. For all these reasons, it is time to embrace the Intelligibility Principle and consign the Nativeness Principle to the past.

Acknowledgements

Many thanks to Pavel Trofimovich for encouraging me to update my 2005 article and for his suggestions on how to make the paper more generally applicable to L2 (rather than English) pronunciation. Many thanks also to Tracey Derwing, Murray Munro and Sinem Sonsaat-Hegelheimer for excellent advice on how to make the arguments clearer.

References

Adank, P., Evans, B.G., Stuart-Smith, J., & Scott, S.K. (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *Journal of Experimental Psychology: Human Perception and Performance*, 35(2), 520–529.

Agudo, J.D.D.M. (Ed.). (2017). *Native and non-native teachers in English language classrooms: Professional challenges and teacher education*. Walter de Gruyter. https://doi.org/10.1515/9781501504143

Baese-Berk, M.M., Bradlow, A.R., & Wright, B.A. (2013). Accent-independent adaptation to foreign accented speech. *The Journal of the Acoustical Society of America*, 133(3), EL174–EL180. https://doi.org/10.1121/1.4789864

Bird, S. (2020). Pronunciation among adult Indigenous language learners: The case of SENĆOŦEN /tʾ/. *Journal of Second Language Pronunciation.* https://doi.org/10.1075/jslp.17042.bir

Bradlow, A.R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, 106(2), 707–729. https://doi.org/10.1016/j.cognition.2007.04.005

Brown, A. (1988). Functional load and the teaching of pronunciation. *TESOL Quarterly*, 22(4), 593–606. https://doi.org/10.2307/3587258

Bohn, O., & Munro, M. (2007). *Language experience in second language speech learning. In honor of James Emil Flege.* John Benjamins. https://doi.org/10.1075/lllt.17

Buckingham, L. (2015). Shades of cosmopolitanism: EFL teachers' perspectives on English accents and pronunciation teaching in the Gulf. *Journal of Multilingual and Multicultural Development*, 36(6), 638–653. https://doi.org/10.1080/01434632.2014.994638

Cerreta, S., & Trofimovich, P. (2018). Engaging the senses: A sensory-based approach to L2 pronunciation instruction for actors. *Journal of Second Language Pronunciation*, 4(1), 46–72. https://doi.org/10.1075/jslp.00003.cer

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.

Clarke, C.M., & Garrett, M.F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America*, 116(6), 3647–3658. https://doi.org/10.1121/1.1815131

Cook, V. (1999). Going beyond the native speaker in language teaching. *TESOL Quarterly*, 33(2), 185–209. https://doi.org/10.2307/3587717

Coppieters, R. (1987). Competence differences between native and near-native speakers. *Language*, 63(3), 544–573. https://doi.org/10.2307/415005

Cutler, A. (1986). Forbear is a homophone: Lexical prosody does not constrain lexical access. *Language and Speech*, 29(3), 201–220. https://doi.org/10.1177/002383098602900302

Derwing, T.M. (2003). What do ESL students say about their accents? *Canadian Modern Language Review*, 59(4), 547–567. https://doi.org/10.3138/cmlr.59.4.547

Derwing, T.M. & Munro, M.J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching and Research*, 42 (4), 476–490. https://doi.org/10.1017/S026144480800551X

Derwing, T.M., & Munro, M.J. (2015). Pronunciation fundamentals. *Evidence-based perspectives for L2 teaching and research*. John Benjamins. https://doi.org/10.1075/lllt.42

Deterding, D. (2012). Pronunciation in World Englishes. *The Encyclopedia of Applied Linguistics*. Retrieved from. https://doi.org/10.1002/9781405198431.wbeal0967

Dragojevic, M., Mastro, D., Giles, H., & Sink, A. (2016). Silencing nonstandard speakers: A content analysis of accent portrayals on American primetime television. *Language in Society*, 45(1), 59–85. https://doi.org/10.1017/S0047404515000743

Flege, J.E., Munro, M.J., & MacKay, I.R. (1995). Factors affecting strength of perceived foreign accent in a second language. *The Journal of the Acoustical Society of America*, 97(5), 3125–3134. https://doi.org/10.1121/1.413041

Gatbonton, E., Trofimovich, P., & Magid, M. (2005). Learners' ethnic group affiliation and L2 pronunciation accuracy: A sociolinguistic investigation. *TESOL Quarterly*, 39(3), 489–511. https://doi.org/10.2307/3588491

Giles, H., Wilson, P., & Conway, A. (1981). Accent and lexical diversity as determinants of impression formation and perceived employment suitability. *Language Sciences*, 3(1), 91–103. https://doi.org/10.1016/S0388-0001(81)80015-0

Gluszek, A., & Dovidio, J.F. (2010). Speaking with a nonnative accent: Perceptions of bias, communication difficulties, and belonging in the United States. *Journal of Language and Social Psychology*, 29(2), 224–234. https://doi.org/10.1177/0261927X09359590

Hahn, L.D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly*, 38(2), 201–223. https://doi.org/10.2307/3588378

Harrison, G. (2013). "Oh, you've got such a strong accent": Language identity intersecting with professional identity in the human services in Australia. *International Migration*, 51(5), 192–204. https://doi.org/10.1111/imig.12005

Harrison, G. (2014). 12 Accent and 'Othering' in the workplace. In J. Levis & A. Moyer (Eds.), *Social dynamics in second language accent* (pp. 255–272). DeGruyter Mouton. https://doi.org/10.1515/9781614511762.255

Jenkins, J. (2000). *The phonology of English as an international language*. Oxford University Press.

Kachru, B.B. (Ed.). (1992). *The other tongue: English across cultures*. University of Illinois Press.

Kang, O., Thomson, R.I., & Moran, M. (2018). Empirical approaches to measuring the intelligibility of different varieties of English in predicting listener comprehension. *Language Learning*, 68(1), 115–146. https://doi.org/10.1111/lang.12270

Kartushina, N., & Frauenfelder, U.H. (2014). On the effects of L2 perception and of individual differences in L1 production on L2 pronunciation. *Frontiers in Psychology*, 5, 1246–1262. https://doi.org/10.3389/fpsyg.2014.01246

LeVelle, K., & Levis, J. (2014). Understanding the impact of social factors on L2 pronunciation: Insights from learners. In J. Levis & A. Moyer (Eds.), *Social dynamics in second language accent* (pp. 97–118). DeGruyter Mouton. https://doi.org/10.1515/9781614511762.97

Levis, J.M. (1999). The intonation and meaning of normal yes/no questions. *World Englishes*, 18(3), 373–380. https://doi.org/10.1111/1467-971X.00150

Levis, J.M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly*, 39(3), 369–377. https://doi.org/10.2307/3588485

Levis, J.M. (2006). Pronunciation and the assessment of spoken language. In R. Hughes (Ed.), *Spoken English, TESOL and applied linguistics* (pp. 245–270). Palgrave Macmillan. https://doi.org/10.1057/9780230584587_11

Levis, J. M. (2018). *Intelligibility, oral communication, and the teaching of pronunciation*. Cambridge University Press. https://doi.org/10.1017/9781108241564

Levis, J. M., & Sonsaat, S. (2017). Pronunciation teaching in the early CLT era. In O. Kang, R. Thomson & J. Murphy (Eds.), *The Routledge handbook of English pronunciation*, (pp. 267–283). Routledge. https://doi.org/10.4324/9781315145006-17

Levis, J. M., Sonsaat, S., Link, S., & Barriuso, T. A. (2016). Native and nonnative teachers of L2 pronunciation: Effects on learner performance. *TESOL Quarterly*, 50(4), 894–931. https://doi.org/10.1002/tesq.272

Lippi-Green, R. (2012). *English with an accent: Language, ideology and discrimination in the United States*. Psychology Press. https://doi.org/10.4324/9780203348802

Mahboob, A., & Golden, R. (2013). Looking for native speakers of English: Discrimination in English language teaching job advertisements. *Age*, 3(18), 21.

Major, R. C., Fitzmaurice, S. F., Bunta, F., & Balasubramanian, C. (2002). The effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL Quarterly*, 36(2), 173–190. https://doi.org/10.2307/3588329

McCrocklin, S., & Link, S. (2016). Accent, identity, and a fear of loss? ESL students' perspectives. *Canadian Modern Language Review*, 72(1), 122–148. https://doi.org/10.3138/cmlr.2582

McCullough, E. A., Clopper, C. G., & Wagner, L. (2019). Regional dialect perception across the lifespan: Identification and discrimination. *Language and Speech*, 62(1), 115–136. https://doi.org/10.1177/0023830917743277

Medgyes, P. (1992). Native or non-native: who's worth more? *ELT Journal*, 46(4), 340–349. https://doi.org/10.1093/elt/46.4.340

Moussu, L., & Llurda, E. (2008). Non-native English-speaking English language teachers: History and research. *Language Teaching*, 41(3), 315–348. https://doi.org/10.1017/S0261444808005028

Moyer, A. (2014). Exceptional outcomes in L2 phonology: The critical factors of learner engagement and self-regulation. *Applied Linguistics*, 35(4), 418–440. https://doi.org/10.1093/applin/amu012

Mugglestone, L. (1995). *Talking proper: The rise of accent as social symbol*. Oxford University Press.

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, 45(1), 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Munro, M. J., & Derwing, T. M. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System*, 34(4), 520–531. https://doi.org/10.1016/j.system.2006.09.004

Munro, M. J., Derwing, T. M., & Sato, K. (2006). Salient accents, covert attitudes: Consciousness-raising for pre-service second language teachers. *Prospect*, 21, 67–79.

Ockey, G. J., & French, R. (2016). From one to multiple accents on a test of L2 listening comprehension. *Applied Linguistics*, 37(5), 693–715. https://doi.org/10.1093/applin/amu060

Pennington, M. C., & Rogerson-Revell, P. (2019). *English pronunciation teaching and research*. London: Palgrave Macmillan. https://doi.org/10.1057/978-1-137-47677-7

Piske, T., MacKay, I. R., & Flege, J. E. (2001). Factors affecting degree of foreign accent in an L2: A review. *Journal of Phonetics*, 29(2), 191–215. https://doi.org/10.1006/jpho.2001.0134

Rubin, D. L. (1992). Nonlanguage factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants. *Research in Higher Education*, 33(4), 511–531. https://doi.org/10.1007/BF00973770

Saito, K., Trofimovich, P., & Isaacs, T. (2016). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics*, 37(2), 217–240. https://doi.org/10.1017/S0142716414000502

Saito, K., Trofimovich, P., & Isaacs, T. (2017). Using listener judgments to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics*, 38(4), 439–462.

Sakai, M. (2018). Moving towards a bilingual baseline in second language phonetic research. *Journal of Second Language Pronunciation*, 4(1), 11–45. https://doi.org/10.1075/jslp.00002.sak

Scharenborg, O. (2007). Reaching over the gap: A review of efforts to link human and automatic speech recognition research. *Speech Communication*, 49(5), 336–347. https://doi.org/10.1016/j.specom.2007.01.009

Smith, L. E., & Rafiqzad, K. (1979). English for cross-cultural communication: The question of intelligibility. *TESOL Quarterly*, 371–380. https://doi.org/10.2307/3585884

Sorace, A. (1993). Incomplete vs. divergent representations of unaccusativity in non-native grammars of Italian. *Second Language Research*, 9(1), 22–47. https://doi.org/10.1177/026765839300900102

Sorace, A. (2003). Near-nativeness. In C. Doughty & M. Long (Eds.), *The handbook of second language acquisition* (pp. 130–151). Wiley Blackwell. https://doi.org/10.1002/9780470756492.ch6

Subtirelu, N. C. (2015). "She does have an accent but…": Race and language ideology in students' evaluations of mathematics instructors on RateMyProfessors.com. *Language in Society*, 44(1), 35–62. https://doi.org/10.1017/S0047404514000736

Suzukida, Y., & Saito, K. (2019). Which segmental features matter for successful L2 comprehensibility? Revisiting and generalizing the pedagogical value of the functional load principle. *Language Teaching Research*, 1362168819858246. https://doi.org/10.1177/1362168819858246

Tan, A. (2020). Transfer of suprasegmental improvements to novel sentences and segmental accuracy using real time audiovisual pitch training. Master's thesis, Iowa State University. https://doi.org/10.31274/etd-20200624-143

Thomson, R. (2014). Accent reduction and pronunciation instruction are the same thing. In L. Grant (Ed.), *Pronunciation myths: Applying second language research to classroom teaching* (pp. 160–187). Ann Arbor: University of Michigan Press.

Weatherhead, D., White, K. S., & Friedman, O. (2016). Where are you from? Preschoolers infer background from accent. *Journal of Experimental Child Psychology*, 143, 171–178. https://doi.org/10.1016/j.jecp.2015.10.011

Expanding the scope of L2 intelligibility research[*]

Intelligibility, comprehensibility, and accentedness in L2 Spanish

Charles L. Nagle and Amanda Huensch
Iowa State University | University of Pittsburgh

This study investigated relationships among intelligibility, comprehensibility, and accentedness in the speech of L2 learners of Spanish who completed a prompted response speaking task. Thirty native Spanish listeners from Spain were recruited through Amazon Mechanical Turk to transcribe and rate extracted utterances, which were also coded for grammatical and phonemic errors, and speaking rate. Descriptively, although most utterances were intelligible, their comprehensibility and accentedness varied substantially. Mixed-effects modeling showed that comprehensibility was significantly associated with intelligibility whereas accentedness was not. Additionally, phonemic and grammatical errors were significant predictors of intelligibility and comprehensibility, but only phonemic errors were significantly related to accentedness. Overall, phonemic errors displayed a stronger negative association with the listener-based dimensions than grammatical errors. These findings suggest that English-speaking learners of Spanish are not as uniformly intelligible and comprehensible as FL instructors might believe and shed light on relationships among speech constructs in an L2 other than English.

Keywords: intelligibility, comprehensibility, accentedness, L2 Spanish

1. Introduction

Twenty-five years ago, Munro and Derwing (1995) demonstrated that comprehensibility and accentedness were distinct, listener-based constructs whose rela-

[*] An earlier version of this article was published as part of a special issue of the *Journal of Second Language Pronunciation* 6(3). https://doi.org/10.1075/jslp.20009.nag

https://doi.org/10.1075/bct.121.04nag

tionship to intelligibility varies across listeners. In their study, comprehensibility was strongly aligned with intelligibility, and within-listener correlations ranged from medium to large. In contrast, the relationship between accentedness and intelligibility was generally weaker and more variable. Since that study, second language (L2) speech research has experienced a theoretical and methodological renaissance centered on the three constructs. For instance, over the past few years, a significant body of scholarship has emerged on the linguistic correlates of comprehensibility and accentedness across multiple speaking tasks (Crowther et al., 2018) and target languages (e.g., Bergeron & Trofimovich, 2017; O'Brien, 2014). Yet, most of this work has concentrated on L2 English, and work that has addressed other L2s has focused on comprehensibility as the primary construct of interest. What is needed, then, is a return to intelligibility, comprehensibility, and accentedness in L2s other than English and in different contexts of learning.

The context of the original studies was English as a Second Language (ESL) in Canada, whereas our focus is on Spanish as a Foreign Language (FL) in the United States. Applying constructs generated in the ESL context to the FL context brings with it a series of conceptual questions related to if and how the constructs need to be adapted. For example, ESL speakers need to be able to communicate with members of the local community so that they can fulfill their immediate needs, which means that local listeners are an appropriate evaluation group. In contrast, FL learners are studying the L2 out of personal and/or professional interest and may not come into contact with proficient L2 speakers other than their instructor during the first few years of FL study. Thus, for FL learners, the question of "Intelligible and comprehensible to whom?" is less straightforward, given that the group of native speakers with whom they might interact is largely imaginary until they study or live abroad. Moreover, FL learners likely envision themselves interacting with a range of native speakers in the US and abroad, which further complicates defining a valid reference group for intelligibility, comprehensibility, and accentedness evaluations. On a more theoretical level, relationships among the constructs may depend on L1-L2 pairings, such that we might expect a slightly different portrait to emerge for L2 Spanish, at least in terms of the magnitude of the attested relationships.

L2 Spanish seems like a logical starting point for expanding the scope of intelligibility, comprehensibility, and accentedness research. Spanish is an important world language. In the US context in particular, it is the most frequently studied FL, both in K-12 (approximately 7.3 million learners, representing 70% of K–12 FL learners; American Councils, 2017) and post-secondary (approximately 1.4 million learners, representing 50% of higher education FL learners; Goldberg et al., 2015) settings. This fact is not surprising since Spanish is the second most spoken language in the US with approximately 38 million speakers

(American Community Survey, 2015). We also find Spanish to be an interesting case since in our experience, many FL Spanish instructors seem to believe that L1 English-speaking learners of Spanish are completely intelligible, and that their intelligibility is not impacted by pronunciation. By investigating intelligibility, comprehensibility, and accentedness in FL learners of Spanish, such claims can be tested and insights into the generalizability of Munro and Derwing (1995) to new L2s and contexts can be gained.

Overall then, revisiting intelligibility, comprehensibility, and accentedness in FL Spanish has the potential (1) to enhance the validity and generalizability of findings by generating parallel evidence in a new research and learning context and by using more sophisticated statistical techniques, which have become widely available in recent years; (2) to begin laying a methodological and conceptual framework for extending intelligibility research to a greater variety of FLs, including less-commonly-taught languages; (3) to shed light on listeners' perception of L2 Spanish speech, which has practical value for FL Spanish instructors and language program directors.

2. Background

In a series of seminal studies, Munro, Derwing, and colleagues (Derwing & Munro, 1997; Munro & Derwing, 1995; Munro et al., 2006) provided evidence of the partial independence of three dimensions of speech: intelligibility (actual understanding of an utterance), comprehensibility (effort required to understand an utterance), and accentedness (the extent to which pronunciation deviates from an expected pattern/norm). Participants in the 1995 study were advanced ESL learners living in Canada and studying at university, whose speech was elicited via a picture description task. Utterances extracted from their narrations were presented to native speakers of English from the local context, who transcribed them and rated their comprehensibility and accentedness. Results indicated that most utterances were transcribed accurately, comprehensibility ratings were somewhat positively skewed, and accentedness ratings were somewhat negatively skewed. Critical findings from that work included evidence that comprehensibility was more related to intelligibility than accentedness and that even some utterances rated as strongly accented were nevertheless transcribed with perfect accuracy. These results provided empirical evidence that being accented was not synonymous with being difficult to understand, and they laid the foundation for a shift in pronunciation research and teaching away from accent reduction toward a focus on comprehensibility and intelligibility (Levis, 2005). To further explore the relationship among these speech dimensions, the authors conducted additional studies with L2 English learners in Canada focusing on the potential impact of speaker L1 (Derwing & Munro, 1997) and listener L1 (Derwing & Munro, 2013; Foote & Trofimovich, 2018; Munro et al., 2006). As in the 1995 study, accentedness, intelligibility, and comprehensibility emerged as partially independent speech dimensions.

Another component of the 1995 and 1997 studies was to investigate the extent to which linguistic features (e.g., phonemic errors, grammatical errors, speech rate) were correlated with the global speech dimensions in an effort to better understand which factors might underlie judgements and/or have an impact on intelligibility. Results indicated that linguistic features were more likely to be related to accentedness/comprehensibility ratings than intelligibility scores, but there was a great deal of interlistener variation in the attested relationships. For instance, in the 1995 study, only 28% of listeners showed significant correlations between phonemic errors and intelligibility, versus 44% and 78% for comprehensibility and accentedness, and there were fewer significant correlations across the board in the 1997 study. Subsequent work examining a greater variety of linguistic predictors has shown that pronunciation and lexicogrammatical features contribute to listener judgments in L2 English (e.g., Crowther et al., 2016; Isaacs & Trofimovich, 2012; Saito et al., 2017; Trofimovich & Isaacs, 2012), German (O'Brien, 2014), French (Bergeron & Trofimovich, 2017), and Japanese (Saito & Akiyama, 2017), but these studies have focused exclusively on comprehensibility and accentedness.

A survey of literature dealing with the FL Spanish context indicates an emphasis on accentedness (i.e., goals of sounding native-like) and a heavy reliance on read speech. One line of inquiry in this area has examined speaker and listener characteristics that affect ratings of foreign accent (e.g., George, 2017; Schoonmaker-Gates, 2015). For example, Schoonmaker-Gates (2015) manipulated the Voice Onset Time (VOT) length of segments in read speech to determine if VOT had an impact on accentedness judgements. Results from her study indicated that both native and nonnative speaker listeners are sensitive to VOT as a marker of foreign accent. Another body of work has examined the extent to which phonetics instruction facilitates gains in pronunciation, as determined by listener ratings or through acoustic comparison of learner productions to a native speaker baseline (e.g., Kissling, 2013; Lord, 2005, 2008). In her survey of Spanish FL instructors, Huensch (2019) observed a tension in instructors' responses, insofar as they seemed to prioritize intelligible speech as an important learning goal while also valuing native-like accuracy (see also Nagle et al., 2018). On the one hand, an emphasis on accentedness in the literature and in the classroom can be important given that more accented speech may be perceived as less grammatical (Ruivivar & Collins, 2018) and may be associated with negative evaluations of intelligence, successfulness, and other markers of social status (Fuertes et al., 2012). On the other hand, an emphasis on accentedness alone has been questioned, for instance, by Kissling (2013) whose conclusion references Derwing and Munro's work and asks "whether accentedness is in fact worthy of future study" (p. 737), arguing that "the most interesting research in the future will balance measures of… accentedness, comprehensibility and intelligibility" (p. 737).

A handful of studies have focused on comprehensibility instead of or in addition to accentedness in the Spanish FL context (e.g., McBride, 2015; Nagle, 2018; Schairer, 1992). For example, Schairer (1992) compared comprehensibility ratings to phonetic analysis of speech samples from English L1 learners of Spanish and concluded that learners' productions of vowels (avoiding reduction to schwa and diphthongization of stressed vowels) best predicted comprehensibility scores. More recently, McBride (2015) had listeners rate speech samples for comprehensibility and pleasantness and additionally asked open-ended questions about what made the samples sound accented or difficult to understand. Issues with fluency and intonation surfaced as the features that had the greatest impact on comprehensibility ratings. Ultimately, little to no FL research has focused on intelligibility either independently or in conjunction with comprehensibility and accentedness. Addressing this gap, the following research questions guided the current study:

Research Questions

1. To what extent are intelligibility, comprehensibility, and accentedness related to one another in beginner L2 Spanish speech?
2. To what extent do linguistic features (i.e., phonemic errors, grammatical errors, speech rate) predict the intelligibility, comprehensibility, and accentedness of beginner L2 Spanish speech?

3. Method

3.1 Participants

3.1.1 *Speakers*

Participants ($n=19$, five men) were recruited from second to fifth semester Spanish courses at a large public university. In their responses to a language background questionnaire (available on IRIS, https://www.iris-database.org/iris/app/home/; Marsden et al., 2016) all participants indicated English as their native language, and when asked about their weekly language use, reported using English a majority of the time: 90–100% ($M=96\%$, $SD=3\%$). Participants had a mean age of 23 ($SD=11$, $range=18–65$) and were majoring in a variety of non-language-related subjects (e.g., Political Science, Biomedical Sciences, Chemistry, Business).

3.1.2 *Listeners*

Following Nagle's (2019) procedure and recommendations, listeners ($n=30$, 23 men) were recruited from Spain using Amazon Mechanical Turk (AMT). Table 1 provides a summary of listener characteristics based on listeners' responses to a language background questionnaire. Listeners self-assessed their proficiency in English and Spanish using 9-point Likert scales (1 = *extremely low proficiency*, 9 = *extremely high proficiency*). On average, they judged themselves to be highly proficient in Spanish and moderately proficient in English, though they reported minimal English use on a daily basis. They also self-evaluated their level of familiarity with nonnative Spanish (1 = *not at all familiar*, 9 = *extremely familiar*), indicating a moderate level of familiarity with non native speech. They reported interacting with nonnative speakers on a monthly or daily basis in both personal and professional contexts. Half had training in linguistics, and a third reported some form of teaching experience. This general listener profile arguably represents the type of listener with whom many FL learners are likely to interact, namely, native listeners who have studied multiple languages and who are reasonably familiar with non native speech.

Table 1. Summary of listener characteristics

	M (SD)	*Range*
Age	31.63 (8.22)	18–48
Age of onset L2 English	6.67 (2.80)	0–12
Global English proficiency	7.00 (1.36)	4–9
Global Spanish proficiency	8.88 (0.31)	8–9
Percent daily English use	13.87 (13.32)	0–50
Familiarity with L2 Spanish	6.33 (2.02)	2–9
Interactions with L2 speakers:	Never: 3 Monthly: 14 Daily: 7 More than daily: 6	
Context of L2 interactions:	Personal: 7 Professional: 7 Both: 14	
Linguistic training:	Yes: 16	
L2 teaching experience:	Yes: 11	

3.2 Materials

Speech data were elicited via a prompted response modeled on the NCSSFL-ACTFL Can-Do Statements: *¿Qué haces en tu tiempo libre?* (What do you do in your free time?). Speaker recordings were transcribed in CLAN following CHAT conventions and checked by the second researcher. Two utterances representing full phrases minus any initial hesitations such as *uh* were selected from each speaker to be used as stimuli for the AMT rating task, for a total of 38 utterances. Utterances ranged between 4–17 words and 2.14–18.63 seconds with a mean length of 9.47 words (*SD*=3.90) and 8.26 seconds (*SD*=4.66). The CLAN transcripts were converted into Praat TextGrids, and segmented TextGrids were used to extract the utterances. The scale peak function in Praat set to 99dB was used to create files of approximately equal loudness for the listening task. Pilot testing with three native speakers indicated that listeners were able to successfully complete the transcription and rating task.

3.3 Procedure

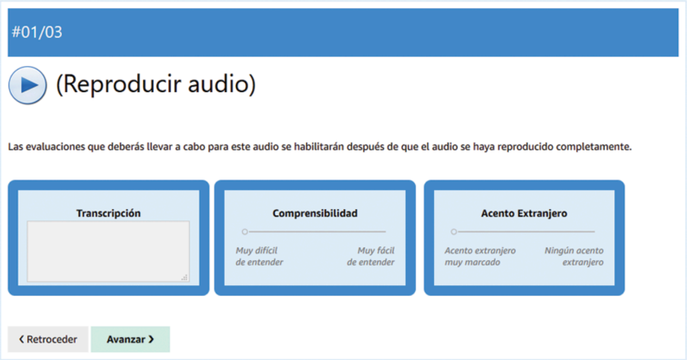
Speaker recording sessions were held individually in a quiet room. After completing the informed consent process, listeners completed a variety of tasks related to a larger project on L2 Spanish learning. For the speaking task used in the current study, participants were instructed to speak for approximately 1 minute in response to the question, *¿qué haces en tu tiempo libre?* Participants were given a few moments to think about their answers before responding. Speakers were compensated with a US$ 20 Amazon gift card.

We used geographic filtering in AMT to recruit online listeners from Spain[1] to transcribe and rate the utterances. After completing a background questionnaire (to be placed on IRIS), listeners were asked to transcribe and rate the 38 utterances presented in a random order while wearing headphones. The task began with instructions and two practice items before continuing to the main task. For each item, listeners pressed play when they were ready to hear the utterance. The task interface required listeners to listen to the complete utterance before having 45 seconds to provide a transcription and their ratings. Listeners were instructed

[1] AMT allows for geographic filtering by country but not by specific regions within countries. Thus, although we attempted to control for dialect influences on ratings using this filtering option, we would like to acknowledge that there are multiple varieties of Spanish spoken within Spain, which is typically divided into two major dialect zones: north/central and southern. We asked participants to indicate the city in which they had been born. Twenty-two listeners were born in central or northern Spain (e.g., Madrid, Segovia, Valencia), six in southern Spain (e.g., Sevilla, Murcia), one in Caracas, Venezuela, and one in Lisbon, Portugal. Although one listener indicated that he was born in Portugal, he nonetheless reported Spanish as one of his L1s.

to write down exactly what they heard and then to rate the comprehensibility and accentedness of each utterance using 100-point sliding scales. Figure 1 is an image of the online AMT rating interface. At the end of the experiment, listeners were asked to rate how well they understood the constructs and the difficulty of the task. They also had the opportunity to provide additional open-ended comments on the task and rating interface. Listeners spent an average of 32 minutes on the task (*SD* = 7.69) and were compensated US$ 4 for their participation, in line with the US federal minimum wage at the time of listener recruitment ($ 7.25/hour).

Preliminary inspection of the transcription and rating results indicated that listeners understood the constructs (on a 100-point scale with 100 being “I understood it very well”, Accentedness, *M* = 91, *SD* = 17; Comprehensibility, *M* = 93, *SD* = 11) and found the task relatively easy to complete (on a 100-point scale with 100 being “Very easy to complete”, *M* = 77, *SD* = 23). The comprehensibility and accentedness data were submitted to reliability analysis using two-way, consistency, average-measure intraclass correlation coefficients (ICC). Results of this analysis indicated excellent reliability for both constructs: for comprehensibility, ICC = .97, 95% CI = [.95, .98] and for accentedness, ICC = .97, 95% CI = [.96, .99].



Las evaluaciones que deberás llevar a cabo para este audio se habilitarán después de que el audio se haya reproducido completamente.

Figure 1. Amazon Mechanical Turk rating interface

3.4 Analysis

3.4.1 *Data coding*

In line with Munro and Derwing (1995), the 38 utterances used in the listening tasks were coded for phonemic and grammatical errors. Phonemic errors were defined as any deletion, insertion, or substitution of a phoneme clearly interpretable as a Spanish phoneme different from the correct one (e.g., [ˈko.ɾo] ‘chorus’ vs. [ˈko.ro] ‘I run’, [ˈmi̯a.ɾo] [no translation] vs. [ˈmi.ɾo] ‘I watch”). Errors in word stress placement and inappropriate vowel reduction were also included (e.g.,

[ˈme.nəs] vs. [ˈme.nos], [ˈbe.ɾe] vs. [be.ˈɾe]). Grammatical errors (e.g., number, gender, preposition use) in each utterance were also counted (e.g., *la*$_{FEM}$ *restaurante*$_{MASC}$, *yo*$_{1stSING}$ *habla*$_{3rdSING}$). Phonemic and grammatical errors were coded by a native Spanish research assistant and checked by the first author. Speech rate was operationalized as the number of syllables per second speaking time (i.e., excluding pauses; this measure has also been referred to as articulation rate and avoids confounds with measures of pausing [De Jong et al., 2013]). To determine utterance length, syllables were counted manually by two coders based on the audio and transcriptions of the 38 utterances. An inter-rater reliability analysis conducted using two-way, agreement, average-measures ICC on the independent coding from two raters on the 38 utterances was high, ICC = 0.99, 95% CI = [.99, .99]. Utterance duration was calculated automatically from the segmented TextGrids (250ms silent pause cutoff, De Jong & Bosker, 2013) using a Praat script.

Transcriptions provided by the listeners were compared to those created by the authors after careful listening and coded for exact word matches. Misspellings (including lack of accent marks, which some listeners did not use) were not considered deviations. Trivial errors such as phonemic and grammatical regularizations (e.g., *telanovela* transcribed as *telenovela* 'soap opera'; *yo habla* transcribed as *yo hablo*) were also coded. Two coders separately completed the coding for 10 of the 30 listeners ($n = 380$ utterances). Inter-rater reliability (two-way, agreement, average-measure ICC) for the exact match (ICC = .99, 95% CI = [.99, .99]) and trivial error (ICC = .91, 95% CI = [.88, .92]) codings was excellent. Therefore, one coder completed the coding of the transcriptions for the remaining 20 listeners. From the coded transcriptions, an intelligibility score was calculated by summing the exact word matches and trivial errors and dividing by the total number of words.

3.4.2 *Mixed-Effect Models*

Mixed-effects models were fit in R version 3.6.1 (R Core Team, 2019) using the lme4 package (Bates et al., 2014). The following covariates were included in all models to control for their relationship with the dependent variables.

- Speaker-level covariates: age of onset L2 Spanish and amount of time learning L2 Spanish.
- Listener-level covariates: age, age of onset L2 English, self-estimated global proficiency in L2 English, daily English use, familiarity with nonnative speech, and previous teaching experience.
- Utterance-level covariates: Number of syllables, mean silent pause duration (computed over the utterance), local speech rate (i.e., articulation rate, com-

puted over the utterance), number of corrections per utterance, and number of repetitions per utterance.

All continuous predictors were z-scored, and for the categorical teaching experience variable, the baseline value was set to zero (i.e., no previous teaching experience). By-speaker and by-listener random effects were fit. All models included random intercepts for both groupings, with additional by-listener random slopes fit for fixed effects of interest, as described below. Likelihood ratio tests were used to compare models and evaluate fit, and QQ plots were used to check the assumption that model residuals were normally distributed. For intelligibility, we opted to fit models to the more lenient intelligibility metric that did not penalize trivial errors.

4. Results

As displayed in Figure 2, most utterances were transcribed with perfect accuracy, comprehensibility ratings were distributed throughout the 100-point scale, and accentedness ratings were skewed toward moderately to strongly accented.

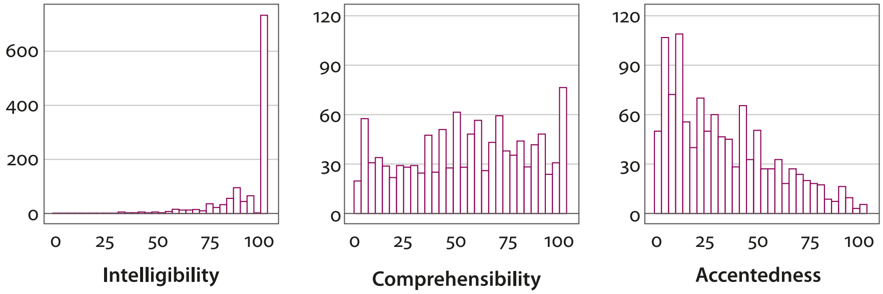


Figure 2. Distribution of intelligibility scores (transformed to a 100-point scale for the sake of display) and comprehensibility and accentedness ratings

4.1 Relationships among intelligibility, comprehensibility, and accentedness

To evaluate the first research question related to relationships among intelligibility, comprehensibility, and accentedness, separate models were fit to the intelligibility and comprehensibility data. Preliminary models fit to the continuous intelligibility variable revealed that residuals significantly deviated from normality. Attempts to bring residuals closer to normality by transforming the data were unsuccessful. Therefore, the continuous measure was recoded into a binary measure where scores $< .90$ were assigned a value of 0 and scores $\geq .90$ were assigned a value of 1. A cutoff of .90 was selected on the basis of previous literature indicating intelligibility rates of .90 to 1 for native speaker utterances. A generalized model, which does not impose the same assumption with respect to normality of model residuals, was then fit to the binary measure.

The primary predictors of interest in this model were the z-transformed comprehensibility and accentedness scores. Generalized models output log-odds, which can be transformed into odds ratios through exponentiation. On the odds ratio scale, a ratio less than 1 indicates that the predictor reduces the probability of an intelligible transcription, whereas a ratio greater than 1 indicates that the predictor enhances the probability of an intelligible transcription. As reported in Table 2, the association between intelligibility and comprehensibility was statistically significant. Utterances that were rated as more comprehensible were far more likely to be transcribed intelligibly. More precisely, an utterance rated as one unit more comprehensible (1 *SD* above the mean) on the z-scored comprehensibility scale would be 3.29 times more likely to be transcribed intelligibly, an utterance with a comprehensibility score of 2 (2 *SD* above the mean) would be 6.58 times more likely to be transcribed intelligibly, and so forth. In contrast to the significant positive association between intelligibility and comprehensibility, the relationship between intelligibility and accentedness missed significance. A number of covariates, however, emerged as significant predictors. With respect to listener-level covariates, listeners who were older on average and who reported more experience with non native speech were more likely to transcribe utterances intelligibly. Finally, with respect to utterance-level covariates, utterances containing a greater-than-average number of repetitions and utterances with a greater-than-average silent pause duration were more likely to be intelligible, whereas longer utterances (i.e., utterances containing a greater-than-average number of syllables) were less likely to be intelligible. Including by-listener random slopes for comprehensibility resulted in a singular fit, suggesting overfit. Therefore, the random effect was not retained. By-listener random slopes for accentedness were not tested since the fixed effect missed significance.

Inspection of the comprehensibility model residuals showed a normal distribution. Thus, the comprehensibility models were fit to the original variable on the 100-point scale. This model contained z-scored intelligibility and accentedness predictors and speaker-, listener-, and utterance-level covariates. Including by-listener random slopes for intelligibility and accentedness significantly enhanced model fit ($\chi^2(5) = 39.22$, $p < .001$), suggesting that there was significant between-listener variation in the strength of the association between both predictors and comprehensibility. As shown in Table 3, there were positive relationships between intelligibility and comprehensibility and between accentedness and comprehensibility. Utterances that were more intelligible and less accented – on the 100-point

Table 2. Summary of generalized mixed-effects model fit to intelligibility scores

Fixed effects	*Odds ratio*	*95% CI*	*p*
Intercept	4.76	[2.53, 8.96]	<.001
Comprehensibility	3.29	[2.53, 4.27]	<.001
Accentedness	.80	[.62, 1.02]	.07
Speaker-level covariates			
Age of onset L2 Spanish	1.49	[.85, 2.62]	.17
Learning time	1.31	[.74, 2.34]	.36
Listener-level covariates			
Age	1.44	[1.09, 1.89]	<.01
Age of onset L2 English	.88	[.67, 1.17]	.39
L2 English proficiency	.85	[.64, 1.13]	.26
Daily English use	1.02	[.78, 1.33]	.91
Familiarity L2 speech	1.37	[1.04, 1.81]	.03
Teaching experience: Yes	1.24	[.68, 2.26]	.48
Utterance-level covariates			
Speech rate	1.29	[.86, 1.92]	.22
Mean silent pause duration	1.35	[1.04, 1.76]	.02
Number of corrections	1.24	[.88, 1.75]	.23
Number of repetitions	1.48	[1.09, 1.99]	.01
Length (syllables)	.56	[.40, .88]	.001
Random effects			
By-speaker intercept	1.28		
By-listener intercept	.25		

Note. All continuous predictors were transformed into z-scores.

scale with higher scores indicating a more targetlike accent – were also more comprehensible. Contrasting with the intelligibility results showing positive effects for pause length and repetitions, the only significant covariate for comprehensibility was self-corrections. Utterances containing a greater-than-average number of self-corrections were rated as less comprehensible. Regarding the by-listener random effects, there was comparatively more variance in the relationship between accentedness and comprehensibility than intelligibility and comprehensibility, as evidenced by the greater *SD* for the former (5.60 for accentedness vs. 3.62 for intelligibility).

Table 3. Summary of mixed-effects model fit to comprehensibility ratings

Fixed effects	*Estimate*	*95% CI*	*p*
Intercept	55.90	[47.58, 64.23]	<.001
Intelligibility	6.95	[5.16, 8.75]	<.001
Accentedness	9.30	[6.66, 11.94]	<.001
Speaker-level covariates			
Age of onset L2 Spanish	−.19	[−5.86, 5.48]	.95
Learning time	1.62	[−4.01, 7.25]	.57
Listener-level covariates			
Age	−3.28	[−8.40, 1.84]	.21
Age of onset L2 English	.61	[−4.50, 5.71]	.82
L2 English proficiency	4.04	[−1.43, 9.50]	.15
Daily English use	.35	[−4.78, 5.47]	.90
Familiarity L2 speech	−3.87	[−9.11, 1.38]	.15
Teaching experience: Yes	−5.13	[−16.06, 5.79]	.36
Utterance-level covariates			
Speech rate	.82	[−1.68, 3.32]	.52
Mean silent pause duration	.84	[−.51, 2.18]	.22
Number of corrections	−3.19	[−5.36, −1.01]	.004
Number of repetitions	1.58	[−.39, 3.55]	.12
Length (syllables)	−.09	[−2.41, 2.23]	.94
Random effects	*SD*		
By-speaker intercept	11.85		
By-listener			
Intercept	13.40		
Intelligibility	3.62		
Accentedness	5.60		

Note. All continuous predictors were transformed into z-scores.

Residuals for the accentedness models were mostly normal, except at the upper end where they were slightly larger than expected. Despite this minor deviation from normality, the distribution of accentedness model residuals was deemed sufficiently normal to proceed with the linear models on the original 100-point accentedness scale. The effects reported in Table 4 confirm findings documented in the intelligibility and comprehensibility models, namely a marginally significant negative relationship with intelligibility – more intelligible utterances were rated as more accented – and a positive relationship with comprehensibility – utterances that were rated as more comprehensible were rated as less accented. With respect to model estimates, there was a far stronger relationship between comprehensibility and accentedness (*estimate* = 8.22) than between intelligibility and accentedness (*estimate* = −1.14). With respect to covariates, utterances spoken at a faster-than-average pace were rated as significantly less accented.

Table 4. Summary of Mixed-Effects Model Fit to Accentedness Ratings

Fixed effects	*Estimate*	*95% CI*	*p*
Intercept	34.56	[27.10, 42.02]	<.001
Intelligibility	−1.14	[−2.21, −.06]	.04
Comprehensibility	8.22	[6.75, 9.70]	<.001
Speaker-level covariates			
Age of onset L2 Spanish	.53	[−3.28, 4.35]	.78
Learning time	1.21	[−2.56, 4.99]	.53
Listener-level covariates			
Age	.63	[−4.93, 6.19]	.82
Age of onset L2 English	.09	[−5.59, 5.78]	.97
L2 English proficiency	−2.98	[−8.81, 2.85]	.32
Daily English use	−.60	[−6.05, 4.86]	.83
Familiarity L2 speech	4.93	[−.72, 10.59]	.09
Teaching experience: Yes	−7.07	[−18.93, 4.80]	.24
Utterance-level covariates			
Speech rate	4.29	[2.16, 6.43]	<.001
Mean silent pause duration	1.04	[−.16, 2.23]	.09
Number of corrections	.69	[−1.22, 2.59]	.48
Number of repetitions	−.86	[−2.59, .88]	.33
Length (syllables)	−.72	[−2.73, 1.28]	.48
Random effects	*SD*		
By-speaker intercept	7.75		
By-listener intercept	13.62		

Note. All continuous predictors were transformed into z-scores.

4.2 Phonemic and grammatical errors

To answer the second research question concerning relationships between phonemic and grammatical errors and intelligibility, comprehensibility, and accentedness, a model was fit to each global speech dimension including the z-scored error variables as predictors as well as their interaction term. Phonemic errors were negatively related to intelligibility (*odds ratio* = .55, 95% CI = [.41, .75], $p < .001$), which shows that utterances containing more errors were less likely to be intelligible. Surprisingly, the relationship between grammatical errors and intelligibility was positive (*odds ratio* = 1.39, 95% CI = [1.07, 1.80], $p = .02$), which would suggest that utterances containing more errors were *more* likely to be intelligible. Because detailed follow-up analyses suggested that this was in fact not the case, we will not discuss this finding further.[2] The phonemic × grammatical errors interaction term was not significant (*odds ratio* = 1.15, 95% CI = [.71, 1.85], $p = .57$). Including by-listener random slopes for the error terms resulted in a singular fit, so those effects were not retained.

Models fit to the comprehensibility and accentedness data included intelligibility as a covariate, which allowed for the estimation of the phonemic and grammatical error predictors while controlling for the overall intelligibility of the utterance. For comprehensibility, utterances containing more phonemic errors were rated as significantly less comprehensible (*estimate* = −4.45, 95% CI = [−6.65, −2.24], $p < .001$), as were utterances containing more grammatical errors (*estimate* = −3.97, 95% CI = [−5.83, −2.11], $p < .001$). As illustrated by the magnitude of the estimates, phonemic errors had a stronger negative effect on comprehensibility than grammatical errors did. As in the intelligibility model, the interaction term failed to reach significance (*estimate* = .37, 95% CI = [−3.05, 3.79], $p = .83$). Including the error terms as by-listener random effects did not significantly improve model fit (for phonemic errors, $\chi^2(1) = .35$, $p = .55$; for grammati-

[2] To probe this finding, we fit a zero-one inflated beta regression model. This type of model is advantageous because it fits a separate model to the inflated values at one, which resolves the problematic residuals in the linear model. At the same time, one principal limitation is that this model, as implemented in the glmmTMB package, only accepts one random effect grouping and thus cannot simultaneously estimate the by-speaker and by-listener random effects in the present study. Thus, we fit two models, one with by-speaker random effects and another with by-listener random effects. In both models, all significant effects from the generalized model remained significant, save grammatical errors. In the by-speaker random effect model, grammatical errors was no longer significant (*estimate* = −.02, *SE* = .05, $p = .70$), and in the by-listener model, it remained significant, but the coefficient was negative (*estimate* = −.11, *SE* = .04, $p = .007$), indicating that utterances containing more grammatical errors were less intelligible, as expected.

cal errors, $\chi^2(2) = .84$, $p = .66$). This suggests that relationships between the error categories and comprehensibility were relatively consistent for the individual listeners sampled in this study. For accentedness, only phonemic errors reached significance (*estimate* = −2.87, 95% CI = [−4.79, −.95], $p = .003$), demonstrating that utterances containing more phonemic errors were rated as more accented. The model containing by-listener random slopes for phonemic errors resulted in a singular fit, so the random effect was not retained.

5. Discussion

5.1 Intelligibility, Comprehensibility, and Accentedness

In the present study, we found a strong, positive association between comprehensibility and intelligibility, a nonsignificant relationship between accentedness and intelligibility, and strong, positive alignment between comprehensibility and accentedness. These results largely fall in line with Munro and Derwing's (1995) original findings, except that whereas they reported a fairly even spread of accentedness scores and comprehensibility scores skewed toward easier to understand, we found the opposite. In our study, accentedness was skewed toward moderately to strongly accented, and comprehensibility scores were distributed throughout the 100-point scale. This difference is likely due to proficiency differences in the two samples: advanced ESL speakers in Munro and Derwing (1995) versus novice to intermediate L2 Spanish learners in our study (see also Derwing & Munro, 1997).

We attempted to test for individual, listener-based variation in relationships among the three constructs through the specification of by-listener random effects. The intelligibility models either did not converge, or they demonstrated a singular fit, which indicates that we were not able to estimate a unique slope for each individual listener in our 30 listener sample. However, the inability to model this variation should not be taken as evidence that it does not exist. In contrast, we were able to incorporate by-listener random slopes for intelligibility and accentedness into the model of comprehensibility. The model-estimated standard deviations for those terms indicated greater variability in the relationship between accentedness and comprehensibility than in the relationship between intelligibility and comprehensibility, reinforcing the view that the latter two constructs are more closely aligned with one another. Thus, in some sense, we were able to replicate using more sophisticated modeling techniques the within-listener correlations that Munro and Derwing (1995) and Derwing and Munro (1997) carried out.

5.2 Phonemic and grammatical errors

Our results diverge somewhat from Munro and Derwing (1995) and Derwing and Munro (1997) with respect to relationships between phonemic and grammatical errors and intelligibility, comprehensibility, and accentedness. Whereas Munro and Derwing (1995) found that most listeners demonstrated significant correlations between both error categories and accentedness, we found no significant relationship between accentedness and grammatical errors. Furthermore, whereas they found that only about 50% of listeners showed significant correlations between the two types of errors and comprehensibility, we found that both types of errors were associated with lower overall comprehensibility and that incorporating by-listener random effects did not enhance model fit, which would suggest that the effect was relatively uniform across the listeners in our sample. Finally, Munro and Derwing (1995) reported relatively few significant correlations between errors and intelligibility (less than 30% for any error type), but we found that phonemic and grammatical errors showed a strong negative relationship with intelligibility. One possible explanation is proficiency differences between the participants in the current study and those in Munro and Derwing (1995). However, our findings also differ from Derwing and Munro (1997), whose speakers more closely resembled our own participants. Overall, they found fewer significant correlations in the 1997 study, but grammar scores showed the strongest relationship to all three constructs, at least in terms of the number of listeners showing a statistically significant correlation. Again, this contrasts somewhat with our finding that phonemic errors were most consistently associated with the listener-based constructs. A final result worth mentioning is that none of the models showed a significant interaction among phonemic and grammatical errors. We intuitively thought that utterances containing more overall errors and more error types would substantially degrade comprehensibility beyond the effects of the individual error categories. However, in the current study that does not seem to be the case. Thus, the relationship between errors and speech dimensions appears to be additive instead of multiplicative.

5.3 Other factors

One of the strengths of the present approach is that through modeling we were able to account for a wide variety of speaker-, listener-, and utterance-based influences on intelligibility, comprehensibility, and accentedness, while also controlling for correlations among the predictors themselves. Typically, researchers focus on variation in one facet (e.g., speakers or listeners), while limiting variation in the others to mitigate potential confounding factors. Though methodologically sound, the reality of communication is that the intelligibility, comprehensibility, or accentedness of any stretch of speech necessarily arises out of the complex interaction of speaker, listener, and stimulus features. Thus, we opted to embrace all three facets of the data, prioritizing phonemic and grammatical errors as predictors while also investigating speaker- and listener-based background variables and utterance-level properties.

Two listener-level covariates were shown to enhance intelligibility: age and familiarity with L2 speech. The effect of age is somewhat surprising and to our knowledge has not been attested in the literature. Perhaps older listeners were more attentive during the task and therefore were able to transcribe utterances more accurately. For now, we leave this as an open question for future research. Our finding that listeners who reported more familiarity with L2 Spanish speech tended to transcribe it more accurately but not rate it as more comprehensible or less accented fits with previous research documenting similar effects (e.g., Kennedy & Trofimovich, 2008). Thus, it seems that familiarity with L2 speech may help listeners understand precisely what the speaker is trying to say, but it does not necessarily reduce processing effort or alter listeners' perceptions of the speaker's accent.

Three utterance-level covariates also emerged as significant predictors of intelligibility: silent pauses, repetitions, and length. Silent pauses and repetitions were positively related to intelligibility, whereas utterance length demonstrated a negative relationship. Intuitively, these findings make sense. Longer pauses and repetitions may have helped listeners sort out precisely what the speaker was saying, boosting intelligibility. In contrast, longer utterances were probably more difficult to remember, and as a result, more difficult to transcribe accurately. Although Munro and Derwing (1995) did not find any significant correlations with utterance length, two methodological differences can account for our significant finding. First, whereas Munro and Derwing (1995) carried out separate correlations between utterance length and the listener-based measures, we integrated utterance length into our models alongside an array of other factors, which arguably allowed us to arrive at more reliable estimates of each individual predictor while controlling for the effects of the other predictors in the models. Second, they defined utterance length as number of words, whereas we operationalized it as number of syllables, which resulted in a greater overall range for the predictor.

Relationships between the covariates and comprehensibility and accentedness were far more limited. Corrections seemed to impair comprehensibility, insofar as utterances containing a greater-than-average number of corrections were rated as less comprehensible. Notably, when phonemic and grammatical errors were entered into the comprehensibility model, the effect of corrections was no longer significant, suggesting that errors may have in fact prompted self-corrections, leading to the observed effect. With respect to accentedness, the only significant covariate was speech rate. Derwing and Munro (1997) reported that 23% of listeners showed significant correlations between speech rate and accent ratings. Previous research also suggests that speech is least accented at rates above 4 syllables per second, at least for English (Munro & Derwing, 2001). In the present study, most utterances were spoken at a slower rate of 3.49 syllables per second (*SD* = .61) excluding pauses, or 2.50 syllables per second (*SD* = .80) with pauses. This could explain why utterances spoken at a faster-than-average pace were rated as less accented in this study.

5.4 Adapting listener-based constructs to a new research context

Working in an ESL context, Munro and Derwing (1995) originally defined intelligibility, comprehensibility, and accentedness in reference to local listeners and local speakers. In other words, the constructs were designed to capture speakers' ability to make themselves understood to a group of listeners with whom they might reasonably interact on a daily basis in their personal and professional lives. Since Munro and Derwing's original work, the constructs have taken on a life of their own and have been applied to different varieties of English (Kang et al., 2018) and different L2s, including German (O'Brien, 2014), French (Bergeron & Trofimovich, 2017), Spanish (Nagle, 2018), and Japanese (Saito & Akiyama, 2016), though most of the L2-other-than-English work has focused on comprehensibility and accentedness. Given how far the constructs have travelled, it seems like the right time to reflect upon any necessary adaptations that might need to take place in order to conduct intelligibility, comprehensibility, and accentedness research in a learning and teaching context that is in many ways radically different from the context in which the constructs were initially defined and measured.

One of the most important issues for FL research is precisely who should evaluate FL learners, since during the first few years of FL study, most, if not all, FL learners will spend a majority of their time interacting with one another and their instructor. In the present study, we opted to recruit online raters from Spain using geographic filtering in AMT. This strategy gave us access to a large pool of potential raters while controlling for some of the variability associated with different dialects of Spanish. Nevertheless, this approach has its limitations. For instance, it is unclear exactly how many participants had been exposed to Peninsular varieties of Spanish, and how many of them would envision themselves interacting with speakers of those varieties in the future. Thus, although the general listener profile could be considered ecologically valid in that many FL learners will likely interact with native listeners who are proficient in multiple languages, somewhat familiar with L2 speech, and interact with L2 speakers in different contexts, there may have been a mismatch between the variety of Spanish that participants had learned and the varieties of Spanish that listeners spoke and with which they were familiar. Due to this potential mismatch, some listeners may have assigned harsher accentedness scores, which could explain why the accentedness data in this study were skewed toward the more accented end of the continuum. A full discussion of methodological choices in rater selection for FL learners is beyond the scope of this paper, but one alternative would be to recruit raters from the dialects to which learners have been exposed through their instructors and course materials, which would ensure greater parity with respect to the FL varieties that speakers and listeners use.

Despite this limitation, the overall score distributions in the present study suggest that listeners found these FL Spanish speakers to be highly intelligible, moderately comprehensible, and moderately to strongly accented. Consequently, though learners were generally intelligible, they were far from uniformly comprehensible, a finding that calls into question the tacit belief that English-speaking learners of Spanish have few intelligibility and comprehensibility issues and that these issues are not related to pronunciation. In fact, phonemic errors were a far stronger predictor than grammatical errors in all three models. Given these findings, it would be advantageous for future research to continue to investigate the intelligibility, comprehensibility, and accentedness of FL speakers of varying proficiency and in various L2s, adopting broader definitions of intelligibility whenever possible. Ultimately, this research can help bridge the gap between ESL and FL pronunciation research while also providing actionable information that can help FL instructors decide what to prioritize in their courses.

Acknowledgements

We would like to thank the participants and our research assistants, especially Aneesa Ali and Bianca Pinkerton

Since this publication, we have continued to investigate intelligibility, comprehensibility, and accentedness in L2 Spanish. See our *Language Learning* Registered Report, Huensch and Nagle (2021). For more information on recruiting online raters through platforms such as Amazon Mechanical Turk, see Nagle and Rehman (2021).

References

Looney, D., & Lusin, N.. (2018). *Enrollments in languages other than English in United States institutions of higher education, summer 2016 and fall 2016: Preliminary report.* https://www.mla.org/content/download/83540/2197676/2016-Enrollments-Short-Report.pdf

American Community Survey. (2015). *Detailed languages spoken at home and ability to speak English for the population 5 years and over: 2009–2013*. https://www.census.gov/data/tables/2013/demo/2009-2013-lang-tables.html

American Councils for International Education. (2017). *The National K-12 Foreign Language Enrollment Survey Report*. https://www.americancouncils.org/sites/default/files/FLE-report-June17.pdf

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). _lme4: Linear mixed-effects models using Eigen and S4_. R package version 1.1.-7. CRAN.R-project.org/package=lme4

Bergeron, A., & Trofimovich, P. (2017). Linguistic dimensions of accentedness and comprehensibility: Exploring task and listener effects in second language French. *Foreign Language Annals*, 50, 547–566. https://doi.org/10.1111/flan.12285

Crowther, D., Trofimovich, P., & Isaacs, T. (2016). Linguistic dimensions of second language accent and comprehensibility. *Journal of Second Language Pronunciation*, 2, 160–182. https://doi.org/10.1075/jslp.2.2.02cro

Crowther, D., Trofimovich, P., Isaacs, T., & Saito, K. (2018). Linguistic dimensions of L2 accentedness and comprehensibility vary across speaking tasks. *Studies in Second Language Acquisition*, 40, 443–457. https://doi.org/10.1017/S027226311700016X

De Jong, N.H., & Bosker, H.R. (2013). Choosing a threshold for silent pauses to measure second language fluency. Paper presented at DiSS, Stockholm.

Derwing, T.M., & Munro, M.J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, 19, 1–16. https://doi.org/10.1017/S0272263197001010

Derwing, T.M., & Munro, M.J. (2013). The development of L2 oral language skills in two L1 groups: A 7-year study. *Language Learning*, 63(2), 163–185. https://doi.org/10.1111/lang.12000

Foote, J.A., & Trofimovich, P. (2018). Is it because of my language background? A study of language background influence on comprehensibility judgments. *Canadian Modern Language Review*, 74, 253–278. https://doi.org/10.3138/cmlr.2017-0011

Fuertes, J.N., Gottdiener, W.H., Martin, H., Gilbert, T.C., & Giles, H. (2012). A metaanalysis of the effects of speakers' accents on interpersonal evaluations. *European Journal of Social Psychology*, 42, 120–133. https://doi.org/10.1002/ejsp.862

George, A. (2017). Effects of listener and speaker characteristics on foreign accent in L2 Spanish. *JSMULA*, 5, 127–148.

Huensch, A. (2019). Pronunciation in foreign language classrooms: Instructors' training, classroom practices, and beliefs. *Language Teaching Research*, 23, 745–764. https://doi.org/10.1177/1362168818767182

Huensch, A., & Nagle, C. (2021). The effect of speaker proficiency on intelligibility, comprehensibility, and accentedness in L2 Spanish: A conceptual replication and extension of Derwing and Munro (1995a). *Language Learning*, 71(3), 626–668. https://doi.org/10.1111/lang.12451

Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition*, 34, 475–505. https://doi.org/10.1017/S0272263112000150

Kang, O., Thomson, R.I., & Moran, M. (2018). Empirical approaches to measuring the intelligibility of different varieties of English in predicting listener comprehension: Measuring intelligibility in varieties of English. *Language Learning*, 68, 115–146. https://doi.org/10.1111/lang.12270 mixed

Kennedy, S., & Trofimovich, P. (2008). Intelligibility, comprehensibility, and accentedness of L2 speech: The role of listener experience and semantic context. *Canadian Modern Language Review*, 64, 459–489. https://doi.org/10.3138/cmlr.64.3.459

Kissling, E.M. (2013). Teaching pronunciation: Is explicit phonetics instruction beneficial for FL learners? *The Modern Language Journal*, 97(3), 720–744. https://doi.org/10.1111/j.1540-4781.2013.12029.x

Levis, J.M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly* 39, 369–377. https://www.jstor.org/stable/3588485. https://doi.org/10.2307/3588485

Lord, G. (2005). (How) can we teach foreign language pronunciation? On the effects of a Spanish phonetics course. *Hispania*, 88, 557–567. https://doi.org/10.2307/20063159

Lord, G. (2008). Podcasting communities and second language pronunciation. *Foreign Language Annals*, 41, 365–379. https://doi.org/10.1111/j.1944-9720.2008.tb03297.x

Marsden, E., Mackey, A., & Plonsky, L. (2016). The IRIS Repository: Advancing research practice and methodology. In A. Mackey & E. Marsden (Eds.), *Advancing methodology and practice: The IRIS repository of instruments for research into second languages* (pp. 1–21). Routledge.

McBride, K. (2015). Which features of Spanish learners' pronunciation most impact listener evaluations? *Hispania*, 98, 14–30. https://doi.org/10.1353/hpn.2015.0001

Munro, M.J., & Derwing, T.M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech: The role of speaking rate. *Studies in Second Language Acquisition*, 23, 451–468. https://doi.org/10.1017/S0272263101004016

Munro, M.J., & Derwing, T.M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, 45, 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Munro, M.J., Derwing, T.M., & Morton, S.L. (2006). The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition*, 28, 111–131. https://doi.org/10.1017/S0272263106060049

Nagle, C. (2018). Motivation, comprehensibility, and accentedness in L2 Spanish: Investigating motivation as a time-varying predictor of pronunciation development. *The Modern Language Journal*, 102, 199–217. https://doi.org/10.1111/modl.12461

Nagle, C. (2019). Developing and validating a methodology for crowdsourcing L2 speech ratings in Amazon Mechanical Turk. *Journal of Second Language Pronunciation*, 5(2), 294–323. https://doi.org/10.1075/jslp.18016.nag

Nagle, C.L., & Rehman, I. (2021). Doing L2 speech research online: Why and how to collect online ratings data. *Studies in Second Language Acquisition*, 43(4), 916–939. https://doi.org/10.1017/S0272263121000292

Nagle, C., Sachs, R., & Zárate-Sández, G. (2018). Exploring the intersection between teachers' beliefs and research findings in pronunciation instruction. *The Modern Language Journal*, 102(3), 512–532. https://doi.org/10.1111/modl.12493

O'Brien, M.G. (2014). L2 learners' assessments of accentedness, fluency, and comprehensibility of native and nonnative German speech: L2 learner assessments. *Language Learning*, 64, 715–748. https://doi.org/10.1111/lang.12082

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. www.R-project.org/

Ruivivar, J., & Collins, L. (2018). Nonnative accent and the perceived grammaticality of spoken grammar forms. *Journal of Second Language Pronunciation*, 5(2), 269–293. https://doi.org/10.1075/jslp.17039.rui

Saito, K., & Akiyama, Y. (2017). Linguistic correlates of comprehensibility in second language Japanese speech. *Journal of Second Language Pronunciation*, 3, 199–217. https://doi.org/10.1075/jslp.3.2.02sai

Saito, K., Trofimovich, P., & Isaacs, T. (2017). Using listener judgments to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics*, 38, 439–462. https://doi.org/10.1093/applin/amv047

Schairer, K. E. (1992). Native speaker reaction to non-native speech. *The Modern Language Journal*, 76, 309–319. https://doi.org/10.1111/j.1540-4781.1992.tb07001.x

Schoonmaker-Gates, E. (2015). On voice-onset time as a cue to foreign accent in Spanish: Native and nonnative perceptions. *Hispania*, 98, 779–791. https://doi.org/10.1353/hpn.2015.0110

Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition*, 15, 905–916. https://doi.org/10.1017/S1366728912000168

Comprehensibility and everyday English use*

An exploration of individual trajectories over time

Beth Zielinski and Elizabeth Pryor
Macquarie University, Australia | Holmesglen Institute, Australia

In this longitudinal study we tracked change in comprehensibility and English use over a 10-month period in 14 L2 English learners (8 beginner, 6 intermediate) settling in Australia. They were interviewed 4 times during the 10 months as part of a larger longitudinal study. English use was reported at each interview using a language map and excerpts from recordings of Interviews 1 and 4 were rated for comprehensibility. Intermediate participants tended to be more comprehensible and maintain a higher level of English use over time than the beginners. Exploration of individual variation revealed a range of comprehensibility outcomes, the variable and non-linear nature of English use trajectories, and a possible relationship between comprehensibility change and English use for some participants. Important methodological implications for future studies relate to the measurement of comprehensibility and English use, the speech samples used for comprehensibility ratings, and the importance of individual variation.

Keywords: longitudinal, comprehensibility, English use, individual variation

1. Introduction

Listener judgements of comprehensibility have been described by Derwing and Munro (2009) as "the gold standard" (p. 478) because they provide insight into how easy or difficult second language (L2)[1] accented speech is to understand from

* An earlier version of this article was published as part of a special issue of the *Journal of Second Language Pronunciation* 6(3). https://doi.org/10.1075/jslp.20011.zie

https://doi.org/10.1075/bct.121.05zie

the listener's perspective. If a listener finds a speaker difficult to understand, effective communication will be compromised. The factors that influence the development of comprehensibility over time are therefore of considerable interest. In the current study, we took a longitudinal perspective to explore the relationship between spoken interactions in English (henceforth referred to as English use) in everyday life and the development of comprehensibility over time. The longitudinal nature of the study allowed us to capture both the general trends in change over time and the individual variation that occurs between beginning and end points. Exploration of individual variation over time provides important insight into the process of development (Verspoor, Lowie, & Dijk, 2008) and identifies the need for more detailed investigation at specific points in time (Larsen-Freeman & Cameron, 2008).

Despite the importance of a longitudinal perspective (see Ortega & Iberri-Shea, 2005), there are very few studies that systematically investigate the development of comprehensibility over time. Although comprehensibility has been used as a measure of improvement over time in some L2 pronunciation intervention studies (see Thomson & Derwing, 2015; Saito & Plonsky, 2019 for reviews), few studies have investigated how comprehensibility develops without targeted instruction (but see Kennedy, Foote, & Buss, 2015; Saito, Dewaele, & Hanzawa, 2017). Likewise, few longitudinal studies have systematically tracked English use in L2 speakers living in an English-speaking environment (but see Ranta & Meckelborg, 2013). Furthermore, to our knowledge, Derwing and Munro and their co-researchers have conducted the only longitudinal study that looks at the relationship between English use and the development of comprehensibility (Derwing & Munro, 2013; Derwing, Munro, & Thomson, 2008). Derwing and Munro's[2] focus was on migrants settling in Canada. Migrants settling in countries such as Canada, Australia, Britain, New Zealand and the US represent a significant group of L2 English learners. We also focused on this group in the current study.

Derwing and Munro's longitudinal study, conducted over a period of 7 years, followed the development of oral language skills and exposure to English in two groups of L2 English learners: L1 Mandarin speakers and L1 Russian and

[1] Although we use the term second language (L2) or English as a second language (ESL) to represent learners of English, we acknowledge that for many of the participants in studies like ours, English is additional to multiple other languages they speak. Similarly, when we refer to a participant's reported first language (L1), we acknowledge that many also reported speaking other languages as well.

[2] For ease of expression, we refer the longitudinal study conducted by Derwing, Munro and colleagues as Derwing and Munro's study. However, we acknowledge the contribution of the other researchers involved.

Ukrainian speakers (referred to as Slavic language speakers). All participants had completed post-secondary education, and at the beginning of the study all were attending full-time, beginner-level English classes. Derwing et al. (2008) reported on the participants' exposure to English (radio, TV and interactions in English of more than 10 minutes) and listener ratings of their comprehensibility and fluency (at the 2 month, 10 month, and 2 year time points). Derwing and Munro (2013) reported on interactions in English of 10 minutes or more and listener ratings of comprehensibility, fluency and accentedness at the 2 month, 2 year, and 7 year time points. Because of the focus of the current study, our discussion here is limited to their findings related to English use and comprehensibility.

A key finding of Derwing and Munro's study was the different developmental trajectories observed for the Mandarin and Slavic language groups in both comprehensibility and English use. Both groups had similar mean comprehensibility ratings at the 2 month time point, but after that the Slavic language group's comprehensibility improved over time, while the Mandarin group showed no significant improvement. Derwing et al. (2008) concluded that the difference between the two groups in comprehensibility development over the first two years may be related in part to their difference in English use, as the Slavic language group reported significantly more English use than the Mandarin group over this time period. However, the relationship between English use and comprehensibility development is not clear in the time period after that. Although the Slavic language group's comprehensibility continued to improve from the 2 year to the 7 year time point, Derwing and Munro (2013) reported very little change in English use in either group over that period of time. They did, however, make some observations about individual comprehensibility outcomes and English use for some Mandarin speakers. When compared numerically, the two participants with the worst comprehensibility ratings at the 7 year time point were both Mandarin speakers who reported using English (interactions >10 minutes) less than once a day at both the 2 year and 7 year time points. In contrast, the two Mandarin speakers with the best comprehensibility ratings reported that they had "extensive interactions in English on a daily basis" (p. 177) at both time points. Unfortunately, no further detail was provided about correspondence between English use and comprehensibility outcomes for other participants from either group.

The relationship between English use and the development of comprehensibility over time is still somewhat unclear. The aim of the current study was to build on the work of Derwing and Munro to investigate this relationship further in a group of migrants learning English in the Australian context. In doing so, we considered the measurement of English use over time, the speech samples used for comprehensibility ratings, comprehensibility at different levels of English proficiency, and the importance of individual variation.

1.1 Measuring English use over time

Derwing and Munro measured English use as participant self reports of the frequency of interactions or conversations with others in English that lasted 10 minutes or more. They argued that shorter interactions or conversations were likely to involve routine daily interactions featuring superficial or formulaic language, and therefore likely to have less impact on the development of comprehensibility. However, they described their measure of English use as "somewhat crude" (Derwing & Munro, 2013, p. 181) and argued that improved, more detailed measures of English use, suited to the lifestyles of their participants, were needed.

The degree to which L2 language learners interact in their L2 has been a topic of interest in the study abroad (SA) context, where a range of contexts of social interaction and language activities have been examined, using instruments such as questionnaires, language logs and social network surveys (see Dewey, 2017 for an overview). However, having been developed for use with university students in various SA programs, such instruments may not be suitable for use with beginner-level L2 English learners settling in an English-speaking country. Derwing and Munro (2013) suggested that an electronic log similar to that developed by Ranta and Meckelborg (2013) might be a feasible option in the context of their participants' lifestyles. Although Ranta and Meckelborg developed the log for use with university students (L1 Mandarin speakers), the detailed and systematic tracking of English use over a period of 6 months highlighted the importance of detailed measurement of both the amount and type of English use at multiple time points, and consideration of individual variation.

1.2 Speech samples used for comprehensibility ratings

The procedure followed for comprehensibility ratings in Derwing and Munro's study is a well-established and accepted practice adopted by numerous researchers since it was first introduced over two decades ago (see Munro & Derwing, 1995). At each time point, the participants were recorded narrating the same picture story, and were given time to familiarise themselves with the pictures before starting. The speech samples presented to the listeners for ratings were created from excerpts taken from the beginning of each recording, with false starts and hesitations at the outset of the recording eliminated, resulting in samples of 20–25 seconds in length. The speech samples obtained at different time points were then randomised for presentation to the listeners. Before rating the samples, the listeners were shown the pictures being described in the samples to control for potential effects of familiarity with the content.

Although acknowledging that consistency across studies allows for comparison of results, Crowther, Trofimovich, Isaacs, and Saito (2015) raised the concern that speech samples from picture-based narrative tasks may not reflect interactions in real world contexts, and may result in findings that are specific to the narrative task. Crowther et al., who investigated the comprehensibility of their participants taking part in sections of speaking tests from the International Language Testing System (IELTS) and the Test of English as a Foreign Language (TOEFL), are among a number of researchers who have included speech samples elicited in more real-life speaking tasks. For example, Derwing, Rossiter, Munro, and Thomson (2004) included recordings of a monologue and a conversation with a researcher in their speech samples; Kennedy et al. (2015) recorded their participants taking part in a mock job interview; Galante and Thomson (2017) included recordings of a retelling of a video story, a role play and a monologue; Cerreta and Trofimovich (2018), whose participants were actors, included speech samples from monologues and scenes from a play; and Crowther (2020) included recordings of spoken interactions between pairs of participants. However, despite the more real-life nature of these speaking tasks, all of the abovementioned studies used only relative short excerpts from each recording for their comprehensibility ratings.

Even though it is accepted that comprehensibility judgements can be made in 20–25 seconds, real-life spoken interactions may last longer. Munro (2018) questioned the extent to which ratings of a short section of a larger sample might represent the general comprehensibility of a speaker. Nagle, Trofimovich, and Bergeron (2019), although investigating L2 Spanish rather than L2 English comprehensibility, adopted a dynamic approach to comprehensibility judgements of longer speech samples (150–290 seconds), allowing their listeners to change their comprehensibility ratings in real time as they listened to each speaker. The listeners then listened again to as much of the speech samples as they needed to in order to provide a single global comprehensibility rating for each speaker. The average time they needed to make this global rating was 40.57 seconds, with a range of 0–153 seconds. Although the listeners were able to provide single comprehensibility ratings for these longer samples, Nagle et al. acknowledged the possibility that the timing and values of the global ratings may have been influenced by familiarity with the speech samples, since the listeners had completed the dynamic ratings first. They also speculated that some listeners may have relied on their memory of the dynamic ratings for the global ratings rather than on the speech itself.

1.3 Comprehensibility and English proficiency

Studies investigating comprehensibility have tended to involve ratings of speech samples from participants with the same levels of English proficiency. The majority of studies have involved participants described as having intermediate or high English proficiency (e.g., Derwing & Munro, 1997; Derwing, Munro & Wiebe, 1998; Munro & Derwing, 1995, 1998; Munro, Derwing & Morton, 2006). High-proficiency participants are also represented in studies involving university students (e.g., Crowther et al., 2015; Isaacs, Trofimovich, & Foote, 2018; Kang, 2010; Kennedy et al., 2015).

Apart from Derwing and Munro's longitudinal study, very few studies investigating comprehensibility have had beginner-level learners as their participants (but see Derwing et al., 2004 and Isaacs & Thomson, 2013), and to our knowledge no studies have explored the relationship between English proficiency levels and comprehensibility. Isaacs, Trofimovich, Yu, and Chereau (2015) investigated the relationship between comprehensibility ratings and the overall IELTS Speaking Band as well as its component scales, but the relationship between comprehensibility and the overall IELTS score (i.e., overall English proficiency) was not discussed. Isaacs, Trofimovich and colleagues (Isaacs & Trofimovich, 2012; Saito, Trofimovich, & Isaacs, 2016; Saito, Webb, Trofimovich, & Isaacs, 2016) have used descriptions such as low, intermediate, and high for their participants, but these relate to comprehensibility ratings rather than English proficiency as used in the current study.

1.4 The importance of individual variation

Derwing and Munro (2013) took the important step of exploring individual comprehensibility outcomes and relating them to English use. They examined numerical changes over time in comprehensibility ratings and reported that 7 out of 11 Mandarin speakers and 2 out of 11 Slavic language speakers became less comprehensible over time. They also reported little change in English use over time, with only 5 of the 22 participants (both groups combined) reporting any increase. As mentioned earlier, Derwing and Munro considered the relationship between English use and comprehensibility development for four of the Mandarin speakers. However, there were no details about how English use might have impacted comprehensibility development for the other participants, or how comprehensibility changed over time for the five participants reported to have had an increase in English use over time (or in fact, which group they were from). Individual trajectories over time for both English use and comprehensibility are therefore crucial to the investigation of the relationship between the two.

1.5 Study design and research question

In this longitudinal study, we used a mixed methods approach to explore the individual variation in the development of comprehensibility and change in English use over time. We investigated the development of comprehensibility over a 10-month period in a group of adult migrants settling in Australia. We aimed to extend the scope of Derwing and Munro's longitudinal study by: (a) using a more detailed measure of English use, (b) using authentic, real-life speech samples for comprehensibility ratings, (c) investigating the relationship between English use and the development of comprehensibility in participants at two different levels of English proficiency, beginner and intermediate, and (d) systematically exploring the individual variation in participant outcomes. We addressed the following research question:

> For beginner- and intermediate-level participants, how do comprehensibility and English use change over time, and how are these changes related?

2. Method

2.1 Participants

The L2 English learners were 14 participants in a larger longitudinal qualitative study designed to follow the language learning progress and early settlement experiences of migrants across Australia as they studied in the Adult Migrant English Program (AMEP) and then afterwards (see Yates et al., 2015). The larger longitudinal study followed two groups of migrants: one group for a period of approximately 1.5 years ($n = 85$) and the other ($n = 60$) for a longer period of approximately 4.5 years. The 14 participants featured in the current study are a subset of the 85 participants who were followed for approximately 1.5 years. Five semi-structured interviews were scheduled during this time: the first four interviews were approximately 3–4 months apart and for most participants spanned a period of approximately 9–10 months from the beginning of the study. The fifth interview was scheduled 6–8 months later. In the current study, we focused on the first four interviews, and refer to them as time points T1, T2, T3, and T4. The participants in the current study were selected from the larger study on the basis of the data they had available for analysis, that is, measures of English use at all four interviews and speech samples of sufficient quality for comprehensibility ratings at T1 and T4.[3]

[3] In the larger longitudinal study, interviews were sometimes missed or delayed due to personal circumstances (e.g., overseas travel, pregnancy and childbirth, illness). Also, it was not always possible to complete all tasks in each interview.

Based on their level of placement for the Certificate of Spoken and Written English (CSWE)[4] at T1, eight of the participants were beginners, equivalent to IELTS overall band score of 1 or 2, and six were intermediate-level, equivalent to IELTS overall band score of 4 (for CSWE/IELTS equivalents see Australian Council of TESOL Associations, 2019). The beginners were all from the same class in an AMEP centre in Melbourne, while the intermediate-level participants were divided equally between two different AMEP centres in Sydney. The beginners (4 males, 4 females) ranged in age from 21 to 65 years and represented five different language backgrounds: Mandarin (4), Somali (1), Japanese (1), Albanian (1) and Kurdish (1). The intermediate-level participants (all females) ranged in age from 23 to 30 years and represented three different language backgrounds: Farsi (2), Korean (2) and Tamil (2). The intermediate-level participants were all educated to at least high school level and four also had tertiary qualifications. The beginners, in contrast, had a range of educational levels: two had tertiary qualifications, three had finished high school only, and three had started but not completed their high school education. Most participants in both groups immigrated to Australia on family or spouse visas. One beginner and two intermediate-level participants were refugees.

2.2 Speech samples

As part of the larger longitudinal study, researchers elicited an extended speech sample from each participant by asking a series of prompt questions during each interview and by keeping comments to a minimum. The participants' answers to two of these were used as speech samples for the comprehensibility ratings at T1 and T4: (1) *What do you like about Australia?* (2) *What don't you like about Australia?* Participants were encouraged to answer the questions in English even if the rest of the interview had been conducted with an interpreter (as was the case with four of the beginners).

Interviews were recorded as part of the larger longitudinal study, using Olympus digital voice recorders, in circumstances beyond our control. Most were conducted in a quiet room, but sometimes it was necessary for researchers to conduct interviews in surroundings that were not ideal for recording purposes (e.g., in a cafe, outside, with children present). If, as judged by us, the recording quality of a speech sample at either T1 or T4 was problematic for comprehensibility ratings to be made, the participant was not included in the current study. Eighteen participants were initially selected, but as a further check for recording quality, we asked the raters to indicate for each sample whether they had problems judging

[4] CSWE levels have since been replaced by a different assessment framework.

the speaker's comprehensibility and, if so, why. An additional four participants were excluded from the study because multiple raters indicated that the recording quality was not good or that there was too much background noise.

The 28 samples used in the current study (14 participants × 2 times) were an average of 48 seconds long and most lasted from 30 to 50 seconds.[5] They were interspersed in a larger set of 41 randomised samples rated as part of another study. As a verification that the raters remained in step while listening to the larger set of speech samples, we included recordings of two native speakers of Australian English answering the two prompt questions as described above (see Derwing & Munro, 2013; Derwing et al., 2008).

2.3 Reported English use

As part of each interview, the participants were guided by the researcher to report their English use in everyday life using a language map. Figure 1 shows English use noted on a language map for Yuan[6] at T4. The language map was developed as part of the larger longitudinal study as a practical way to structure the participants' reported estimates of their use of English in a range of different contexts and then to elicit an estimation of their overall English use, that is, the percentage of time they used English in everyday life. Figure 1 shows the part of the language map related to English use. The full language map also includes information about languages used when reading and writing, and when watching TV or listening to the radio. The participants in the larger study had a range of English language skills and educational backgrounds, and some had little education. The language map was therefore designed to be used face to face at each interview, using an interpreter where necessary.

2.4 Raters

Ten raters, all with tertiary qualifications, were recruited from the research team involved in the larger longitudinal study. They represented a mix of language backgrounds: three were highly proficient nonnative speakers of English (NNSs) from German, Mandarin and Ga L1 backgrounds, and seven were native English speakers, two of whom also spoke another language (Tamil, Italian). Our rationale for including a mix of language backgrounds stems in part from the multicultural

[5] Two longer speech samples were inadvertently included in the rated samples. We decided to include the ratings in our analysis but have discussed the possible implications in the Discussion.

[6] All names used are pseudonyms.

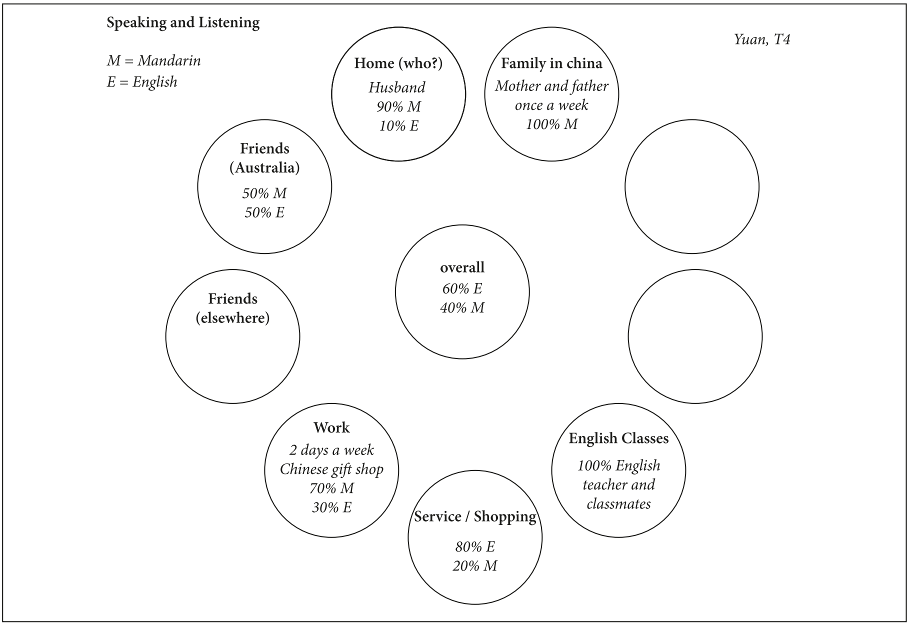


Figure 1. Language map: Yuan, T4

and multilingual nature of the cities in which the participants lived, Melbourne and Sydney (Chik, Benson & Moloney, 2019; The State of Victoria, Department of Premier and Cabinet, 2018). Furthermore, Derwing and Munro (2013) found that native speakers and highly proficient NNSs rate the comprehensibility of L2 accented speech in a similar way. The raters also had a range of expertise in the area of teaching and assessing L2 speech. Four of the native English speakers and two of the highly proficient NNSs had training and experience in teaching ESL; the four others did not.

2.5 Rating procedure

Raters completed the rating task individually, sourcing the electronically provided speech samples on their computers and listening to them wearing headphones. One week before the main rating task, we provided the raters with five practice speech samples to familiarise them with the procedure. The practice samples included four other participants from the larger study and one native speaker, all answering the same prompt questions featured in the main rating task samples.

The raters were provided with written instructions for the practice and main rating tasks, which included the number of speech samples they would hear, the topic the people in the samples would be talking about (i.e., what they like, and don't like about living in Australia), and how long most of the samples would be. They were provided with electronic versions of the forms on which to record their ratings and had the option of completing the rating electronically or in hard copy. The samples were presented in the same order for all raters. As part of another study, the samples were rated for both comprehensibility and fluency, but only the comprehensibility ratings were used for the current study. The raters were instructed to rate comprehensibility the first time they listened to the samples, and to either take a break before rating fluency, or complete the fluency ratings on a separate occasion.

The scale used for comprehensibility ratings was a 5-point scale (1 = *extremely difficult to understand*, 5 = *very easy to understand*). This scale is different in length to those used in Derwing and Munro's longitudinal study; Derwing et al. (2008) used a 7-point scale and Derwing and Munro (2013) a 9-point scale. Our decision to opt for a 5-point scale originated in our own difficulty reaching consensus using a 9-point scale when trialling the rating procedure on similar speech samples for a different study. In making this choice we considered the findings of Isaacs and Thomson (2013), who found that although there were pros and cons for each scale length, mean scores obtained for comprehensibility did not differ when 5- or 9-point scales were used. We also reversed the order of the scale used by Derwing and Munro, in line with the 5-point scale used by Isaacs and Thomson. Thus, in the current study the higher the comprehensibility rating, the more comprehensible the speaker.

Raters were instructed to listen to the whole sample before making a decision, use the whole scale when rating, and to consider how difficult or easy it is to understand both the words and the meaning of what the person is saying. For each sample, raters also had the option to indicate if they had a problem rating the speaker and to comment on why.

2.6 Interrater reliability

As was the case in Derwing and Munro's longitudinal study, all raters recognised the native speaker samples and rated them as 5, indicating that they had not lost their place during the task. The native speaker ratings were not included in any further analyses. The raters demonstrated a high level of agreement in their comprehensibility ratings according to the Cronbach's alpha analysis ($\alpha = 0.93$). We therefore averaged the ratings for each sample to derive a mean comprehensibility score at T1 and T4 for each participant.

3. Results

In presenting the findings we focus on individual variation in the development of comprehensibility and English use over time. Because of the small sample size and the nature of the data, differences in individual outcomes and possible patterns in group data are discussed as tentative observations using descriptive information only.

3.1 Change in comprehensibility over time

Individual comprehensibility outcomes for beginner- and intermediate-level participants are presented in Table 1. Included is each participant's comprehensibility score at T1 and T4, their change in comprehensibility over time, and the average comprehensibility scores for each group. As shown, across both groups, there was some numerical change in comprehensibility scores for all but one intermediate-level participant (Iris), and no two participants had the same T1 to T4 trajectory. Rezarta, a beginner, had the best outcome (+1.1) and Takumi, another beginner, had the worst (−0.9). As a group, at both T1 and T4, the intermediate-level participants tended to be more comprehensible than the beginners. Average comprehensibility scores for the intermediate-level group were higher than the beginners at both time points, and this trend is also apparent in Figure 2, where we have plotted individual comprehensibility scores at T1 and T4 for each group.

As shown in Figure 2 (and detailed in Table 1), most intermediate-level participants had comprehensibility scores of 4 or above at both time points, while all of the beginners had comprehensibility scores of 3 or below at T1, and this was also the case for most of them at T4. Also shown, the intermediate-level participants tended to have less variability in their comprehensibility scores than the beginners at both time points. The beginners had a wider range of comprehensibility scores than the intermediate-level participants at T1, and the range widened even further for the beginners at T4 (T1, Beg: 1.7–3.0; Int: 3.7–4.4. T4, Beg: 1.5–3.9; Int: 3.8–4.4). Furthermore, it can be seen from Table 1 and Figure 2 that the range of individual change in comprehensibility scores over time was wider for the beginners than for the intermediate-level group. For the beginners, the change ranged from an increase of 1.1 (Rezarta) to a decrease of 0.9 (Takumi), while for the intermediate-level group the changes were more modest, ranging from an increase of 0.3 (Rose and Karen) to a decrease of 0.5 (Nina).

Table 1. Change in comprehensibility over time for beginner- and intermediate-level participants

			Comprehensibility scores		
Name	**Age**	**L1**	**T1**	**T4**	**Numerical change**
			Beginner		
Rezarta	21	Albanian	2.8	3.9	+1.1
Adam	27	Somali	2.2	3.1	+0.9
Yuan	35	Mandarin	2.7	3.3	+0.6
Shan	43	Mandarin	2.6	2.8	+0.2
Nas	40	Kurdish	1.7	1.5	−0.2
Ying	47	Mandarin	2.5	2.3	−0.2
Liam	32	Mandarin	2.7	2.5	−0.2
Takumi	65	Japanese	3.0	2.1	−0.9
Average (range)			**2.5 (1.7–3.0)**	**2.7 (1.5–3.9)**	**−0.2 (−0.9–1.1)**
			Intermediate		
Rose	25	Farsi	4.1	4.4	+0.3
Karen	23	Farsi	4.0	4.3	+0.3
Chellam	30	Tamil	3.7	3.9	+0.2
Iris	26	Korean	4.2	4.2	+0.0
Mathu	28	Tamil	4.4	4.2	−0.2
Nina	27	Korean	4.3	3.8	−0.5
Average (range)			**4.1 (3.7–4.4)**	**4.1 (3.8–4.4)**	**0.0 (−0.5–0.3)**

Note. Participants in each group are arranged in descending order of the numerical change in comprehensibility over time (higher comprehensibility scores correspond to improved comprehensibility).

3.2 Change in English use over time

Individual English use trajectories for beginner- and intermediate-level participants are presented in Table 2. Included is the overall English use reported by each participant at each time point, the average overall English use for each group at each time point, and each participant's average overall English use across all time points. As shown, there were various changes in overall English use over time for all but one participant (Shan), and these changes were not necessarily linear in nature. Furthermore, no two participants had the same English use trajectory.

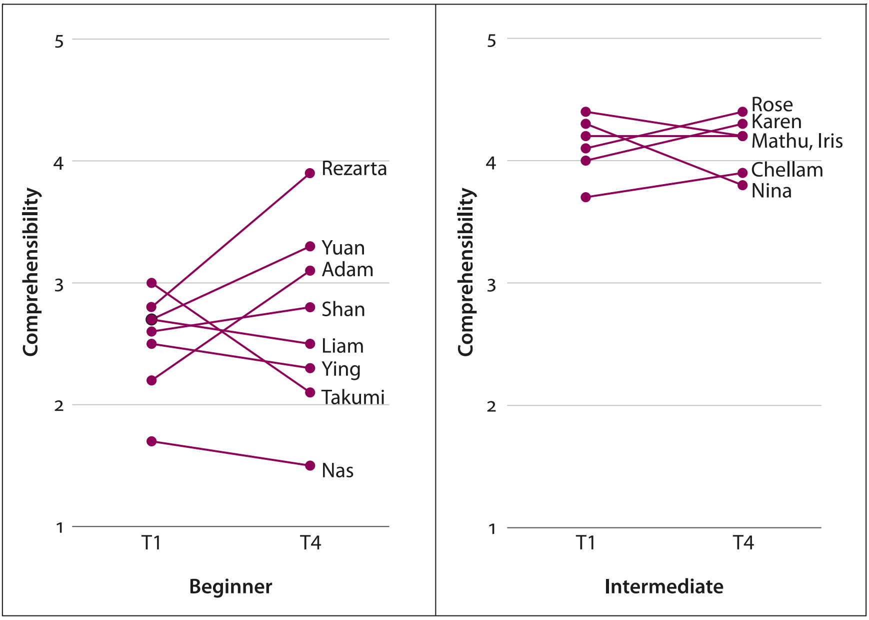


Figure 2. Comprehensibility scores at T1 and T4 for beginner- and intermediate-level participants

Notwithstanding the variation in English use trajectories, as a group, the intermediate-level participants tended to use English more than the beginners throughout the time period. Group averages plotted in Figure 3 show the intermediate-level group's higher average overall English use at all four time points. Also evident is a similar pattern in both groups of an increase in English use at T2 followed by a decrease at T3.

Individual intermediate-level participants also seemed to maintain their level of English use more consistently over time, showing less variability across time points than the beginners. As shown in Table 2, apart from Rose, who had a substantial increase in English use from T1 (20%) to T2 (85%) and then further fluctuations at T3 (60%) and T4 (90%), most changes across time points for the intermediate-level participants were 10% or less. In addition, all but one (Iris) maintained English use of 50% or more from T2 onwards. In contrast, the beginners showed more changes in their English use across the time points. Apart from Shan, who reported the same English use at each time point, all beginners had at least one increase or decrease of more than 10% in their trajectory. Furthermore, only two beginners (Shan and Yuan) maintained English use of more than 50% consistently across multiple time points, and three reported using close to no English at one time point (Takumi, 5% at T1; Rezarta, 5% at T4; Ying, 2% at T3).

Table 2. English use trajectories for beginner- and intermediate-level participants

	Overall English use (% of time English is used)				
Name (age)	**T1**	**T2**	**T3**	**T4**	**Average**
		Beginner			
Shan (43)	80	80	80	80	80.0
Yuan (35)	60	60	90	60	67.5
Adam (27)	40	40	25	40	36.3
Liam (32)	10	50	30	30	30.0
Nas (40)	20	33	10	30	23.3
Takumi (65)	5	30	15	20	17.5
Rezarta (21)	20	30	10	5	16.3
Ying (47)	30	10	2	10	13.0
Average (range)	**33.1 (5–80)**	**41.6 (10–80)**	**32.8 (2–90)**	**34.4 (5–80)**	**35.5**
		Intermediate			
Nina (27)	80	90	80	90	85.0
Chellam (30)	70	80	80	80	77.5
Rose (25)	20	85	60	90	63.8
Karen (23)	50	50	60	50	52.5
Mathu (28)	30	50	50	50	45.0
Iris (26)	40	50	40	40	42.5
Average (range)	**48.3 (20–80)**	**67.5 (50–90)**	**61.7 (40–80)**	**66.7 (40–90)**	**61.0**

Note. Participants in each group are arranged in descending order of their average overall English use.

3.3 The relationship between change in comprehensibility and English use

At first glance the relationship between change in comprehensibility and English use seems somewhat tenuous. On the one hand, as shown in Tables 1 and 2, Rezarta, the participant with the greatest gains in comprehensibility over time, had an average overall English use of only 16.3%, and reported using close to no English (5%) at T4. On the other hand, Nina, who had the highest average overall English use across both groups (85.0%), became less comprehensible over time. However, Rezarta and Nina aside, there is some indication of a relationship between average overall English use and change in comprehensibility over time

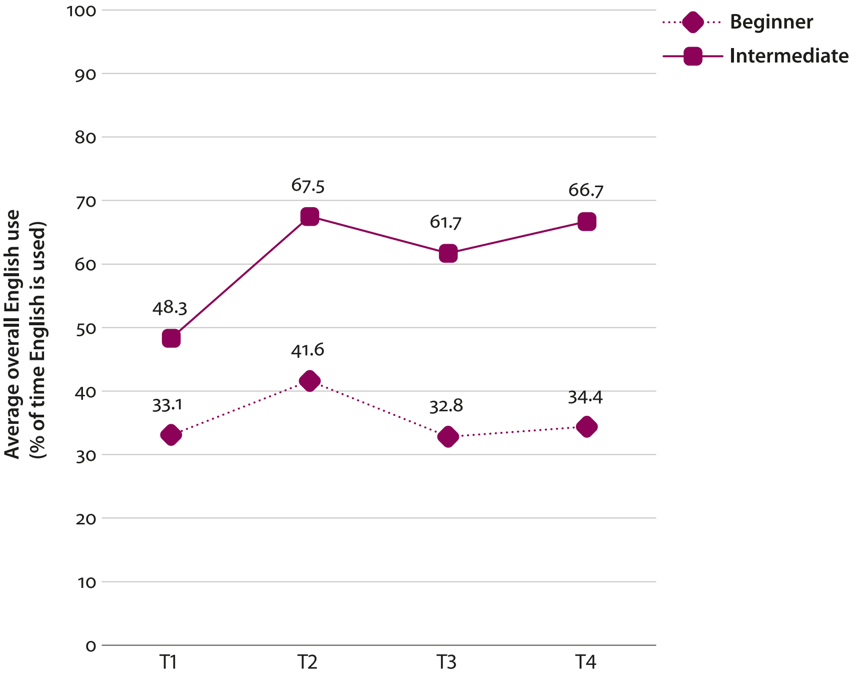


Figure 3. Average overall English use trajectories for beginner- and intermediate-level participants

for participants in both groups, that is, those participants in each group whose comprehensibility scores increased over time tended to be those in their group with the highest average overall English use. As shown in Tables 1 and 2, the three beginners (excluding Rezarta) whose comprehensibility scores increased, Adam (+0.9), Yuan (+0.6), and Shan (+0.2), were in the top three in their group for average overall English use. Similarly, the three intermediate-level participants whose comprehensibility scores increased, Rose (+0.3), Karen (+0.3), and Chellam (+0.2), were also in the top three (excluding Nina) for average overall English use in their group.

3.4 Contexts of English use

Analysis of the contexts in which the participants reported using English suggests that for some participants, it may not be the average overall English use, but the contexts in which English is used that is important for improvement in comprehensibility. The participants reported using English in a range of contexts, and these changed over time. Contexts of English use included home, public services (e.g., shopping, library, bank, doctor), places of worship, the AMEP (with class members during class and in breaks), further education facilities, work, and socialising with friends.

The contexts in which two intermediate-level participants, Rose and Nina, used English provide some insight into the impact of expanding (or not expanding) contexts of English use on English use trajectories and potentially their different comprehensibility outcomes. Both Rose and Nina had similar attributes at T1. Although they were from different language backgrounds (Farsi and Korean respectively), both were intermediate-level, a similar age at T1 (25 and 27 years respectively), with a similar level of education (two years of college following secondary school), and both had started their AMEP classes within a month of each other.

As shown in Tables 1 and 2, Nina's average overall English (85.0%) was higher than Rose's (63.8%), but Rose's comprehensibility score increased over time (+0.3) while Nina's decreased (−0.5). Rose and Nina's reported overall English use trajectories are presented in Figure 4, which shows both participants ended up using English 90% of the time at T4. However, while there was little change in Nina's English use over time, Rose's changed considerably in a non-linear trajectory from 20% at T1 to 90% at T4.

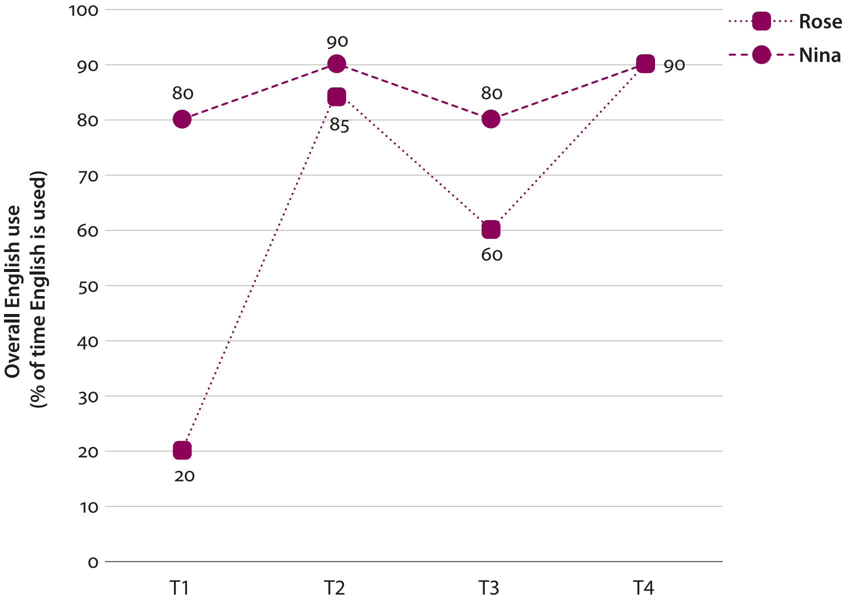


Figure 4. Rose and Nina: Overall English use reported at each time point

Table 3. Rose (R) and Nina (N). Contexts of English use reported at each interview

		Interview			
Context		**T1**	**T2**	**T3**	**T4**
Home	R	Lives with her parents and speaks Farsi at home.			
	N	Uses English at home with her husband (an Australian who speaks only English) and his parents.			
Friends	R	Uses a small amount of English with fellow students at the AMEP. No English-speaking friends elsewhere.	Has more friends at the AMEP, so uses English much more with them now.	Uses English with most of her friends. Speaks English with a very close friend whose L1 is Vietnamese.	Most of her friends are from different L1 backgrounds, so she uses English with them.
	N	Uses a small amount of English with fellow students at the AMEP. No English-speaking friends elsewhere.	Uses English with fellow students at the AMEP and at work experience. All friends elsewhere are Korean.	No English with friends. All friends are Korean.	No English with friends. All friends are Korean.
Formal English classes or further education	R	AMEP	AMEP	AMEP Customer Service Pathways to Employment course with work experience.	Advanced English for Further Studies at TAFE college.
	N	AMEP	AMEP Customer Service Pathways to Employment course with work experience.	Certificate III Information Technology at TAFE college.	Certificate III Information Technology at TAFE college.

Note. In Australia, TAFE colleges are providers of mainly vocational courses. See https://www.tafensw.edu.au/about for information. For information about AMEP Customer Service Pathways to Employment courses see https://www.tafensw.edu.au/student-services/adult-migrant-english-program-amep/pep.

Table 3 presents a summary of the contexts in which Rose and Nina each used English most. Included are English use with family and friends, and the different educational pathways they pursued after finishing their AMEP classes. As shown, Nina used English at home, at the AMEP, and while studying Information Technology at a technical and further education (TAFE) college. However, although Nina's use of English at home contributed to her consistently high overall English use across all time points, she revealed at T4 that even though she always used English with her husband, she did not actually speak to him much. She commented:

> When he finish job we meet in the home, we talking how was your job and how was your day? … We're talking one hour and later on we watching movie, just watching … we're not talking a lot … we're just comfortable with not talking.

Furthermore, although Nina used English when attending AMEP classes, her main friends there and elsewhere spoke Korean. Also, at TAFE college, she rarely spoke to anyone in her class and described her classmates as young males who were native speakers of English. In addition, the nature of the course required the class to spend a lot of time in front of computers completing online activities.

In contrast, Rose did not use English at home, but her friendships with fellow students from different language backgrounds continued to develop over time, thus increasing her English use. At T2 she seemed to be consciously increasing her English use and commented, "I have some Iranian friends, but I prefer to speak English". At T4 she was enrolled in an advanced English course at TAFE college, where she used English for more advanced activities (e.g., oral presentations, discussions about current affairs) and made more English-speaking friends. Thus, at T4 both Nina and Rose were using English 90% of the time, but Nina's English throughout was mostly at home or school (AMEP and TAFE), while Rose's contexts of use expanded over time, from using English with a few of her classmates at T1 to socialising with a large circle of friends and studying advanced English at T4.

Although Rose's increasing circle of friends had an impact on her English use trajectory, other participants reported having no English-speaking friends throughout the study (e.g. Ying), and some only used English with classmates when attending classes (e.g., Nina), or saw their English-speaking friends very occasionally (e.g., Iris, Shan). Takumi's only English-speaking friends outside of the AMEP were part of his church community. He attended meetings with them twice weekly but reported having only simple repetitive conversations with them.

Working rather than studying also had the potential to impact English use trajectories. For Yuan, working had a negative impact on her English use. At T3 Yuan was attending full-time English classes, but at T4 she had changed to part-time classes in order to work in a Chinese gift shop where she spoke mainly Mandarin. Similarly, life circumstances had the potential to impact some participants' English use trajectories in a negative way. For example, at both T3 and T4 Ying had recently returned from trips to China. She had therefore spent little time in

Australia between T2 and T4 and had suspended her English classes. Similarly, Rezarta's circumstances changed after T2. Rather than continuing with her AMEP classes, at T3 she spent most of her time at home with her Albanian-speaking parents-in-law because she was pregnant. At T4 she was at home with a young baby and used very little English.

4. Discussion

In this longitudinal study, spanning a period of 10 months, and focusing on individual variation, we aimed to investigate the change in comprehensibility and overall English use over time, and consider the relationship between the two. We focused on L2 English learners at two different proficiency levels: beginner and intermediate. Our findings confirm the importance of longitudinal studies in exploring individual trajectories over time and raise several key methodological issues.

There was some numerical change in comprehensibility scores over the 10-month time period for all but one intermediate-level participant and no two participants had the same T1 to T4 comprehensibility trajectory. This finding aligns with the variation of individual comprehensibility outcomes observed by Derwing and Munro (2013) between the 2-year and 7-year time points, even though we covered a much shorter time period (10 months vs. 5 years). Intermediate-level participants tended to be more comprehensible than beginners at both the beginning and end of the time period. Most intermediate-level participants had comprehensibility scores of 4 or above at both time points, while most beginners had scores of 3 or below. Intermediate participants also showed less variability in their comprehensibility development over time. The beginners' change over time ranged from +1.1 to −0.9, while for the intermediate-level learners the range was from +0.3 to −0.5.

Most participants reported multiple changes in their overall English use over the four time points, and no two participants had the same English use trajectory. Overall English use did not increase in a linear way over time, and there was considerable variation observed in some participants' trajectories. When compared to the beginners, the intermediate-level participants tended to maintain a higher level of English more consistently across time and showed less variation in their trajectories. Most intermediate-level participants maintained overall English use of 50% or more across multiple time points, while this was the case for only two beginners. Furthermore, three beginners reported using close to no English at one time in their trajectory. These findings complement the observation made by Ranta and Meckelborg (2013), that living in an English-speaking environment does not necessarily ensure that total English exposure increases steadily over time. They also support the importance of Ranta and Meckelborg's systematic and detailed approach to measuring English use over time.

The notion of a relationship between overall English use and the development of comprehensibility was challenged by two participants: Rezarta, who had the greatest gain in comprehensibility but used very little English at T3 and T4, and Nina, who had maintained an overall English use of 80% or more across all time points but became less comprehensible over time. However, there was some indication of a relationship between the comprehensibility and English use for the other participants; those in each group whose comprehensibility scores increased over time were those in their group with the highest average overall English use. This trend complements the findings reported by Derwing et al. (2008) that improvement in comprehensibility over time may be related in part to how much English the participants use, although we covered a shorter time period (10 months vs. 2 years).

4.1 Comprehensibility ratings

Scales of various lengths have been used by researchers for comprehensibility ratings but there is no general agreement as to what scale length is best (Munro, 2018). Munro affirmed the use of a 9-point scale for rating comprehensibility for research purposes, but also noted the possibility of using longer, seemingly continuous scales with only end points marked (see for example, Saito, Trofimovich, & Isaacs, 2017). It is important that the scale used provides raters with enough scope to rate the comprehensibility of the speech samples presented. In our study we had both beginner- and intermediate-level participants and although our decision to use a 5-point scale may have provided the raters with enough range to rate the beginners effectively, it seems that this may not have been the case for the intermediate-level participants. Thus, the finding that intermediate-level participants showed less variability than the beginners in their comprehensibility development over time may have been related to the constraints of the scale rather than a difference in outcomes. As shown in Table 1, most intermediate participants had comprehensibility scores of 4 or above at both time points. Examination of the individual ratings for each participant revealed that all had more than one rating of 5 at T1. This means that for those raters who gave a participant a 5 at T1, there was no way for the scale to accommodate an improvement at T4. It seems, therefore, that the 5-point scale may not have revealed differences for the intermediate participants that might have been evident had a 9-point scale been used.

Another possible influence on the findings of our study was the unsupervised rating of speech samples. Munro and Derwing (2015) advised against unsupervised rating of speech samples because of the lack of control over the conditions in which the samples are rated and the extent to which the raters follow the instructions. The effects of possible differences in rating procedure followed by each rater in our study are therefore unknown.

4.2 Measuring English use

Derwing and Munro (2013) argued that more fine-grained measures of English use are needed. Although the language map used in the current study does provide more detail, it is not without its own limitations. It was developed to provide a structure within which each participant could report not only their overall English use, but also their English use in different contexts. However, like many other measures of English use, it relies on retrospective self-reporting, which can be unreliable (Saito & Akiyama, 2017; Ranta & Meckelborg, 2013). Furthermore, the percentages reported in different contexts were sometimes misleading. For example, a participant who reports using English 100% of the time with her husband may not actually speak to him much, as was the case for Nina (see Section 3.4). Similarly, a participant who reports using English 100% of the time with a certain group of friends, might only see those friends very occasionally, or one who reports using English 100% of the time at work might use simple repetitive English or may not be able to speak much while working because of ambient noise or workplace rules. Sometimes the participants were questioned further about their English use in the different contexts (e.g., *How often do you see that friend? What do you talk about? Who do you talk to at work?*), but this was not consistent across all interviews conducted in the larger longitudinal study, so the relevant information was not available for all participants.

Ranta and Meckelborg (2013) argued for suitable measurement tools to capture not only the amount, but the type of English use at multiple time points. Despite its shortcomings the language map provided us with a means to explore the different contexts in which the participants used English over the time period and highlighted how English use in different contexts might influence English use trajectories and possibly comprehensibility outcomes. However, further research is needed to refine the language map and to determine whether it is an accurate measure of English use for both research and classroom purposes. The question of how the different contexts contribute to overall English use also needs to be addressed, as does the relative importance of different contexts of English use for the development of comprehensibility.

4.3 The relationship between English proficiency and comprehensibility

The finding that intermediate participants tended to be more comprehensible than beginner participants at both T1 and T4 raises the question of the relationship between English proficiency and comprehensibility. Further research is needed to explore the relationship between English proficiency and comprehensibility and whether the linguistic dimensions that influence listeners' comprehensibility ratings (Isaacs & Trofimovich, 2012; Saito, et al., 2017; Trofimovich & Isaacs, 2012) vary for speakers at different proficiency levels. Such research is important for teachers wanting to improve comprehensibility in L2 English learners at different proficiency levels.

4.4 Speech samples

Although the longitudinal nature of the current study provided us with rich information about individual trajectories, we were reliant on data collected as part of the larger longitudinal study. The recordings of the interviews were not intended to be used for the purposes of collecting quality sound files for comprehensibility judgements. The interviews took place in a variety of environments and were conducted by different members of the research team. Some were of good quality, but we had to exclude some potential participants because of the poor sound quality of their recordings at either T1 or T4. Interestingly, Isaacs et al. (2015) experienced a similar frustration with their reliance on sound files obtained from secondary sources rather than recorded in ideal circumstances for the purposes of their study.

In ideal situations, recordings would be made in a soundproof room with high-quality recording equipment. However, in future research we need to be aware of the balance between the naturalness of the speech sample (i.e., interviews in the current study and live IELTS speaking tests in Isaacs et al., 2015) and the need for good sound quality, and all that this entails. In the larger longitudinal study, the researchers were keen to accommodate the participants' life circumstances and commitments and interviewed them when and where was most convenient. Frustratingly, this resulted in a number of speech samples that we could not use in the current study. Future research needs to investigate the impact of sound quality on judgements of comprehensibility in order to inform researchers wanting to collect speech samples in realistic contexts.

On average, our samples were 48 seconds long and most lasted from 30 to 50 seconds, but, as mentioned in Section 2.2, two longer speech samples were inadvertently included. Two of the raters commented that the longer of these samples (Nina, Int, T4, 220 seconds) was too long compared to the others, and one of these raters also made the same comment about the other longer sample (Mathu, Int, T4, 141 seconds). Inclusion of these longer samples has potential repercussions on the intermediate-level participants' outcomes, since both were T4 recordings of intermediate-level participants. However, none of the other raters made mention of the length difference, so we are unsure of the impact on our findings.

Further research is needed to look at the impact of longer speech samples on comprehensibility outcomes so we can interpret research using different sample lengths. Such research is important if we want to investigate comprehensibility in speakers participating in longer interactions in real-life contexts. Munro (2018) suggested that scale size suitability might be contingent on the type of speech sample being rated. Further research is needed to establish whether scales of different lengths or dynamic ratings such as those used by Nagle et al. (2019) might be more suitable for different types and lengths of speech samples.

5. Conclusion

The longitudinal perspective of this study provided insight into group trends and individual variation in small groups of beginner- and intermediate-level participants. Our participant numbers were small and our observations therefore tentative and descriptive in nature. Although there was some indication of a relationship between comprehensibility and English use, a larger study might provide more definitive findings. In order to be able to explore this relationship further, the measurements of both comprehensibility and English use need to be both accurate and valid (Dewey, 2017). Our exploration of individual variation has raised questions about these measurements that point to areas for further investigation in future research. These include the need for more fine-grained measures of English use, the use of longer real-life speech samples for comprehensibility ratings, and the use of different scales and procedures to rate the comprehensibility of different types of speech samples.

Increased understanding of the measurement of change in both comprehensibility and English use over time and the relationship between the two is not only important for researchers, but also ultimately teachers and the learners in their classrooms who benefit from empirically based instruction and advice, and who are the motivation for our interest in this topic.

Acknowledgements

The longitudinal study from which our data is drawn was funded by the Australian Government departments with responsibility for the AMEP. We would like to thank the participants and our colleagues on the project team for their involvement in collecting the data. The views expressed in this article are ours and do not necessarily represent the views of either the government departments or our colleagues.

References

Australian Council of TESOL Associations. (2019). *Submission to the evaluation of the Adult Migrant English Program (AMEP) "new business model"*. Retrieved from: https://tesol.org.au

Cerreta, S., & Trofimovich, P. (2018). Engaging the senses: A sensory-based approach to L2 pronunciation instruction for actors. *Journal of Second Language Pronunciation*, 4(1), 46–72. https://doi.org/10.1075/jslp.00003.cer

Chick, A., Benson, P., & Moloney, R. (Eds.). (2019). *Multilingual Sydney*. London: Routledge.

Crowther, D. (2020). Rating L2 speaker comprehensibility on monologic vs. interactive tasks. What is the effect of speaking task type? *Journal of Second Language Pronunciation*, 6(1), 96–121. https://doi.org/10.1075/jslp.19019.cro

Crowther, D., Trofimovich, P., Isaacs, T., & Saito, K. (2015). Does a speaking task affect second language comprehensibility? *The Modern Language Journal*, 99(1), 80–95. https://doi.org/10.1111/modl.12185

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, 19(1), 1–16. https://doi.org/10.1017/S0272263197001010

Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching*, 42(4), 476–490. https://doi.org/10.1017/S026144480800551X

Derwing, T. M., & Munro, M. J. (2013). The development of L2 oral language skills in two L1 groups: A 7-year study. *Language Learning*, 63(2), 163–185. https://doi.org/10.1111/lang.12000

Derwing, T. M., Munro, M. J., & Thomson, R. I. (2008). A longitudinal study of ESL learners' fluency and comprehensibility development. *Applied Linguistics*, 29(3), 359–380. https://doi.org/10.1093/applin/amm041

Derwing, T. M., Munro, M. J., & Wiebe, G. (1998). Evidence in favor of a broad framework for pronunciation instruction. *Language Learning*, 48(3), 393–410. https://doi.org/10.1111/0023-8333.00047

Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning*, 54(4), 655–679. https://doi.org/10.1111/j.1467-9922.2004.00282.x

Dewey, D. P. (2017). Measuring social interaction during study abroad: Quantitative methods and challenges. *System*, 71, 49–59. https://doi.org/10.1016/j.system.2017.09.026

Galante, A., & Thomson, R.I. (2017). The effectiveness of drama as an instructional approach for the development of second language oral fluency, comprehensibility, and accentedness. *TESOL Quarterly*, 51(1), 115–142. https://doi.org/10.1002/tesq.290

Isaacs, T., & Thomson, R.I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, 10(2), 135–159. https://doi.org/10.1080/15434303.2013.769545

Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition*, 34(3), 475–505. https://doi.org/10.1017/S0272263112000150

Isaacs, T., Trofimovich, P., & Foote, J.A. (2018). Developing a user-oriented second language comprehensibility scale for English-medium universities. *Language Testing*, 35(2), 193–216. https://doi.org/10.1177/0265532217703433

Isaacs, T., Trofimovich, P., Yu, G., & Chereau, B.M. (2015). Examining the linguistic aspects of speech that most efficiently discriminate between upper levels of the revised IELTS pronunciation scale. *IELTS Research Report Series*, 4. Retrieved from https://www.ielts.org/teaching-and-research/research-reports

Kang, O. (2010). Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System*, 38(2), 301–315. https://doi.org/10.1016/j.system.2010.01.005

Kennedy, S., Foote, J.A., & Buss, L.K. (2015). Second language speakers at university: Longitudinal development and rater behaviour. *TESOL Quarterly*, 49(1), 199–209. https://doi.org/10.1002/tesq.212

Larsen-Freeman, D., & Cameron, L. (2008). Research methodology on language development from a complex systems perspective. *The Modern Language Journal*, 92(2), 200–213. https://doi.org/10.1111/j.1540-4781.2008.00714.x

Munro, M.J. (2018). Dimensions of pronunciation. In O. Kang, R.I. Thomson, & J.M. Murphy (Eds.), *The Routledge handbook of contemporary English pronunciation*. (pp. 413–431). New York: Routledge.

Munro, M.J., & Derwing, T.M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, 45(1), 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Munro, M.J., & Derwing, T.M. (1998). The effects of speaking rate on listener evaluations of native and foreign-accented speech. *Language Learning*, 48(2), 159–182. https://doi.org/10.1111/1467-9922.00038

Munro, M.J., & Derwing, T.M. (2015). A prospectus for pronunciation research in the 21st century. A point of view. *Journal of Second Language Pronunciation*, 1(1), 11–42. https://doi.org/10.1075/jslp.1.1.01mun

Munro, M.J., Derwing, T.M., & Morton, S.L. (2006). The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition*, 28(1), 111–131. https://doi.org/10.1017/S0272263106060049

Nagle, C., Trofimovich, P., & Bergeron, A. (2019). Toward a dynamic view of second language comprehensibility. *Studies in Second Language Acquisition*, 41(4), 647–672. https://doi.org/10.1017/S0272263119000044

Ortega, L., & Iberri-Shea, G. (2005). Longitudinal research in second language acquisition: Recent trends and future directions. *Annual Review of Applied Linguistics*, 25, 26–45. https://doi.org/10.1017/S0267190505000024

Ranta, L., & Meckelborg, A. (2013). How much exposure to English do international graduate students really get? Measuring language use in a naturalistic setting. *Canadian Modern Language Review*, 69(1), 1–33. https://doi.org/10.3138/cmlr.987

Saito, K., & Akiyama, Y. (2017). Video-based interaction, negotiation for comprehensibility, and second language speech learning: A longitudinal study. *Language Learning*, 67(1), 43–74. https://doi.org/10.1111/lang.12184

Saito, K., Dewaele, J. M., & Hanzawa, K. (2017). A longitudinal investigation of the relationship between motivation and late second language speech learning in classroom settings. *Language and Speech*, 60(4), 614–632. https://doi.org/10.1177/0023830916687793

Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning*, 69(3), 652–708. https://doi.org/10.1111/lang.12345

Saito, K., Trofimovich, P., & Isaacs, T. (2016). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics*, 37(2), 217–240. https://doi.org/10.1017/S0142716414000502

Saito, K., Trofimovich, P., & Isaacs, T. (2017). Using listener judgments to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics*, 38(4), 439–462. https://doi.org/10.1093/applin/amv047

Saito, K., Webb, S., Trofimovich, P., & Isaacs, T. (2016). Lexical profiles of comprehensible second language speech. The role of appropriateness, fluency, variation, sophistication, abstractness, and sense relations. *Studies in Second Language Acquisition*, 38(4), 677–701. https://doi.org/10.1017/S0272263115000297

The State of Victoria Department of Premier and Cabinet. (2018). *Population diversity in Victoria: 2016 census. Local government areas.* Retrieved from: https://www.multicultural.vic.gov.au

Thomson, R. I., & Derwing, T. M. (2015). The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics*, 36(3), 326–344. https://doi.org/10.1093/applin/amu076

Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition*, 15(4), 905–916. https://doi.org/10.1017/S1366728912000168

Verspoor, M., Lowie, W., & Van Dijk, M. (2008). Variability in second language development from a dynamic systems perspective. *The Modern Language Journal*, 92(2), 214–231. https://doi.org/10.1111/j.1540-4781.2008.00715.x

Yates, L., Terraschke, A., Zielinski, B., Pryor, E., Wang, J., Major, G., … Williams Tetteh, V. (2015). *Adult Migrant English Program (AMEP) Longitudinal Study 2011–2014: Final report.* Sydney: Department of Linguistics, Macquarie University.

Long-term effects of intensive instruction on fluency, comprehensibility and accentedness[*]

Leif M. French, Nancy Gagné and Laura Collins
Sam Houston State University | Université TÉLUQ | Concordia University

We assessed the long-term effects of intensive instruction on different aspects of L2 oral production. Adopting the tridimensional model of oral production (Munro & Derwing, 1995a), we compared high school learners who had received intensive ESL instruction ($N=42$) with non-intensive learners ($N=39$) on perceptual measures of L2 fluency, comprehensibility, and accentedness 4 years after a 5-month intensive instruction period. After controlling for academic ability and L2 proficiency, listeners' ratings of fluency and comprehensibility were significantly higher for the IG; however, there was no specific group advantage for accentedness, suggesting both groups exhibited similar L2 accents. This study provides new empirical evidence that the oral fluency and comprehensibility benefits of an intensive experience may be long-lasting, even when learners' subsequent classroom exposure to the language is much more limited.

Keywords: long-term effects, intensive instruction, second language, oral fluency, fluency, comprehensibility, accentedness, development of oral competence, children

1. Introduction

It is well documented that concentrating the hours of L2 instruction rather than dispersing them over long periods of time (e.g., *Massed* versus *drip-feed* instruction; Collins, Lightbown, Halter, & Spada, 1999; Serrano, 2012) can produce considerable progress in different aspects of L2 development over a relatively short

[*] An earlier version of this article was published as part of a special issue of the *Journal of Second Language Pronunciation* 6(3). https://doi.org/10.1075/jslp.20026.fre

https://doi.org/10.1075/bct.121.06fre

period in both children and adults (Collins & White, 2012; Freed, Segalowitz, & Dewey, 2004; Lightbown & Spada, 1994; Muñoz, 2012; Serrano & Muñoz, 2007; White & Turner, 2005). However, one important question to arise from this body of research is whether L2 learning advantages associated with intensive instruction are maintained over time once learners return to the more limited exposure conditions typical of many foreign language instruction contexts (Collins & Muñoz, 2016). In fact, most studies on intensive instruction have tested L2 learning retention immediately after the intensive program making it difficult to assess the long-term effects because of the short retention interval (Serrano, 2012). Moreover, although research on oral proficiency in intensive programs has focused on different dimensions of oral production such as accuracy (Mora & Valls-Ferrer, 2012) and communicative effectiveness (Collins & White, 2011), the perception of learners' speech by competent speakers of the target language has received little attention.

The current study adopts the widely-documented tridimensional model of oral production (Derwing & Munro, 2013; Derwing, Munro, & Thomson, 2008; Derwing, Munro, Thomson, & Rossiter, 2009; Derwing, Rossiter, Munro, & Thomson, 2004; Munro & Derwing, 1995a). To examine this issue, we compared for the first time high school intensive EFL learners with non-intensive learners on perceptual measures of L2 fluency, comprehensibility, and accentedness 4 years after a 5-month intensive instruction period. The students, matched on academic ability and overall L2 proficiency, were Grade 10 French speakers from the same high school in Quebec, Canada. Throughout high school, all had followed the same French academic program, had maintained a minimum grade average of 60%, and had received the same type and amount of EFL instruction (~four hours a week). The main distinguishing factor was that one group had participated in a 5-month intensive English program in Grade 6, whereas the other had only received regular drip-feed English instruction. Both groups completed a picture-cue narrative task, and expert raters scored the resulting speech samples to determine whether the intensive English group demonstrated any performance advantage with respect to the three perceptive dimensions of oral production 4 years after their program.

1.1 Intensive instruction

Intensive instruction most often refers to a learning context in which the hours of instruction are concentrated into specific blocks of time, providing the opportunity for extensive practice in the L2 often for several hours daily (Collins & White, 2011; Muñoz, 2012). The length of experience and intensity of exposure can vary widely. However, despite this variation, studies in a variety of contexts have generally shown that, for both children and adults, concentrating the hours of L2 instruction rather than spreading them thinly over long periods of time can lead to substantial progress in overall L2 learning, including different aspects of L2 fluency development (Collins & White, 2012; Freed et al., 2004; Huensch & Tracy-Ventura, 2017a, 2017b; Lightbown & Spada, 1994; Muñoz, 2012; White & Turner, 2005).

One common classroom-based approach for designing an intensive learning experience is to target both the amount of instructional time and its distribution. For example, in Quebec, Canada, intensive programs with young learners, also known simply as "Intensive English," increase instruction time and provide a more concentrated distribution of the time over the school year (Lightbown, 2014). In these programs, learners typically receive significantly more exposure and practice than in regular programs (e.g. 300–400 hours vs 60 hours over the school year). The intensive learners are also exposed to a strong version of communicative language teaching several hours a day in which pedagogical focus is overwhelmingly on meaning rather than the specific linguistic features of the language. Therefore, the instructional context provides learners with numerous opportunities for intensive and frequent L2 practice in a variety of communicative situations. As such, in a relatively short time, most learners experience remarkable growth in listening and reading comprehension (Collins & White, 2011; Lightbown & Spada, 1989) accuracy and fluency (Collins & White, 2011; Lightbown & Spada, 1989; White & Turner, 2005), and communication skills (Collins, Lightbown, Halter, & Spada, 1999).

Interestingly, although intensive programs generally show that learners make significant progress in different aspects of oral performance (see Collins & White, 2011; Lightbown & Spada, 1991; White & Turner, 2005), studies so far have relied almost exclusively on objective measures of oral ability (volume of oral production, rate of turn-taking, extent of communicative effectiveness, etc.) during the performance on paired or individual tasks (e.g., information gap, role plays, interviews). Less studied is the more subjective dimension of the perception of learners' speech by competent speakers of the target language. In fact, although the goal of most intensive instruction approaches in L2 teaching is to develop overall oral proficiency in the target language (Lightbown, 2014), there has been little consideration of how such proficiency might be interpreted outside the classroom by competent speakers in the general community. It may be that ability measured by objective tests of oral proficiency largely reflects language professionals' conceptualization of the ability to communicate in the L2 (Sato & McNamara, 2018). Intensive programs would therefore clearly benefit from further research adopting a tridimensional model of oral production, such as that proposed by Derwing and her colleagues (Derwing & Munro, 2013; Derwing et al., 2008; Derwing et al., 2009; Derwing et al., 2004; Munro & Derwing, 1995a, 1995b), to examine subjective or perceived dimensions of fluency (listeners' perception of the smoothness or flow of the speaker's language output), comprehensibility (ease or difficulty with which a listener understands L2 accented speech) and accentedness (degree to which listeners perceive the presence of a foreign accent). Adopting such a model would not only contribute to a better understanding of the perceptual dimensions of intensive learners' oral production, it would also enable us to examine how such dimensions evolve over time outside the intensive setting, which is of particular interest in the present context.

1.2 Long-term effects of intensive instruction

The existing evidence clearly points to the benefits of intensive instruction on different aspects of L2 learning. However, the repeated success of intensive instruction also leaves open the question of whether it leads to significant long-term advantages in L2 learning, particularly when learners return to regular L2 instruction programs. Most research in this area has measured learning outcomes immediately after the intensive program without subsequent post-testing. It may, therefore, be that the observed L2 learning gains at the end of an intensive program result in part from a testing recency effect (Serrano, 2012), making it virtually impossible to determine to what extent these gains are maintained over longer periods of time.

Ideally, to examine the long-term effect of intensive instruction on L2 learning, it is necessary to implement a longitudinal design to follow the same group of learners over time as they transition back to more traditional forms of L2 instruction. There are only a handful of studies that have examined the long-term effects of intensive L2 exposure longitudinally (e.g., Huensch & Tracy-Ventura, 2017a, Llanes, 2012; Regan, 2005), but these have mainly targeted study abroad experiences, which differ significantly from classroom intensive programs in terms of the type, duration, and intensity of language exposure. Nevertheless, findings from these studies suggest that learning gains (both written and oral performance), as a result of intensive L2 exposure during study abroad, are largely maintained up to 15 months later. One recent longitudinal study (Huensch, Tracy-Ventura, Bridges, & Cuesta Medina, 2019) further suggests that L2 oral fluency gains (in particular speech rate) can be preserved up to 4 years after the study abroad experience.

Although study abroad research has increasingly established a link between the intensity of L2 exposure and the extent to which resulting L2 gains are conserved over time, there has been virtually no research that has examined the long-term effects of intensive instruction programs on later L2 learning. In one of the few studies, Lightbown and Spada (1991) compared the ESL performance of two groups of Grade 11 high school students, one of which had previously participated in an intensive ESL program in Grades 5 or 6. They found that intensive English learners performed significantly better than their age peers on oral tasks assessing fluency and accuracy, and were more likely to seek out opportunities to use English outside of school. The learning advantage was maintained even after returning to a regimen of drip-feed ESL instruction throughout secondary school, suggesting that benefits associated with the intensive program 5 years earlier had persisted over time.

The findings reported in Lightbown and Spada (1991) are indeed encouraging with respect to the long-term benefits of intensive classroom instruction. However, there is a need for additional research in these settings to address potential confounding factors when evaluating the long-term effects of intensive instruction. In particular, to better isolate specific long-term effects, it would be important to compare learners with similar intensive learning experiences and academic abilities. For example, in Lightbown and Spada (1991), groups were not previously compared on L2 skills, making it difficult to determine whether intensive learners may have started their program with an L2 learning advantage. Furthermore, the intensive learner group came from different programs across Quebec, leaving open the possibility of potential teaching effects as a result of differences in content and intensity of input across programs. An additional unknown factor was the degree to which schools supported L2 use outside the intensive classrooms, creating, in turn, a potential school effect. More importantly, intensive programs differ greatly in their selection criteria, often relying on academic performance. This suggests the possibility that learners' academic ability at least partially mediated L2 learning outcomes. Additional research in intensive settings, controlling for such factors, would further our understanding of the long-term effects of intensive instruction.

Thus far, there is convincing evidence that learning advantages associated with intensive programs may persist over time, indicating positive long-term instructional effects. However, it remains unclear to what extent other factors besides instruction may moderate these advantages. In the present study, we therefore controlled for potentially confounding factors (age, L2 proficiency, intensive learning experience, academic ability) and investigated the long-term effects of intensive instruction on three specific dimensions of French-speaking high school students' oral production (perceived fluency, comprehensibility and accentedness) 4 years after one group had completed a 5-month intensive English program in elementary school. Specifically, we asked the following question: After controlling for specific intervening variables, is the perceived fluency, comprehensibility and accentedness of IG learners (as rated by expert speakers of English) superior to that of RG learners?

2. Method

2.1 Participants and study context

Participants in this study were Grade 10 French-speaking students (*N* = 81; aged 15–16 years) recruited voluntarily from a public secondary school located in a predominately francophone region of central Quebec (i.e., 85% French-speaking; *Institut de la statistique du Québec*, 2011), which affords students limited opportunities to interact in English. At the time of the study, all were enrolled in an International Baccalaureate Program (IBP) designed to foster intercultural connections between their studies and the real world, and they were required to maintain a minimum grade average of 60% in their core academic subjects – French and Math. All began learning English at the age of 8 (Grade 3) in a limited-exposure ESL program set out by the Quebec Ministry of Education. We excluded from our results students who had one or more parents who was a native speaker (NS) of English and those who had previous extensive contact with English outside the school curriculum, either in summer language programs or repeated travel to English-speaking communities in Canada and the USA.

Table 1. Distribution of ESL instruction (Total number of hours of instruction at the elementary and secondary levels): Intensive and regular groups

	Elementary					**Secondary**					
	Grade 3	**Grade 4**	**Grade 5**	**Grade 6**	**Total Elem.**	**Sec. 1**	**Sec. 2**	**Sec. 3**	**Sec. 4**	**Total Sec.**	**Total**
IG	20	20	60	400	**500**	100	100	100	100	**400**	**900**
RG	20	20	60	60	**160**	100	100	100	100	**400**	**560**

Note. Regular (RG; *n* = 39); Intensive (IG; *n* = 42); Sec. = Secondary.

We recruited three intact classroom groups in Grade 10 (approximately 90 students) from which 81 students ultimately agreed to participate in the study. Of these, 42 had participated in an Intensive English Program in Grade 6 (aged 11–12) with the same experienced, highly proficient ESL teachers. In this program, the intensive group (IG) received full days of theme-based English instruction spread over a 5-month period (~400 hours), specifically targeting the development of speaking and listening skills. The remaining cohort, which constituted the regular group (RG; $N=39$), received no intensive instruction. Instead, the RG was enrolled in Quebec's core ESL program, where they received regular drip-feed English instruction (approximately two hours weekly) dispersed throughout elementary school. By the end of elementary school (i.e., Grade 6), the IG and RG had received approximately 500 and 160 total hours of formal English instruction, respectively (See Table 1).

Upon entry to high school and the International Program, both groups participated in the same core ESL curriculum and received the same type and amount of task-based instruction (~three hours weekly) with the same ESL instructors over a 4-year period. In other words, their experience with English in high school was the same: 400 hours of instruction. At the time of data collection, the IG and RG had received approximately 900 and 560 total hours, respectively, of English instruction (see Table 1). Thus, the main distinguishing factor between these groups was that the IG had a greater number of L2 instruction hours, with a period of concentrated L2 instruction over a relatively short time period in Grade 6.

Further, over the first four years of secondary school, participants in both groups maintained a minimum GPA of 60% for core academic subjects (Math and French) and for ESL (assessed by school board exams targeting students' ability to interact orally as well as understand and produce texts). However, for the purpose of this study, we verified whether the IG and RG actually differed with respect to academic performance and overall English ability to control for the potential confounds of both academic ability and L2 proficiency on observed oral production outcomes. Consequently, we compared both groups' end-of-term percentile scores for mathematics, French and English in Grade 10 (Secondary 4) using Mann-Whitney U tests and then calculated r-values to estimate effect sizes (Tomczak & Tomczak, 2014). These tests revealed that both groups were indeed statistically similar for mathematics ($U=689.5$, $p=221$; $r=.136$), French ($U=717.5$, $p=.336$, $r=.106$) and English ($U=739$, $p=449$, $r=.084$). The effect sizes were also relatively small, suggesting further that at the time of the study both groups demonstrated similar academic abilities and proficiency in English (see Figure 1).

To summarize, the essential distinguishing characteristics of the context and groups examined in this study were the following: (1) since both groups lived and went to school in a predominately French-speaking region, they had little opportunity for daily contact with English outside the classroom and, as such, in this context, English could clearly be considered a foreign language rather than a L2; (2) both study groups started ESL in a limited-exposure program in Grade 3; but, in Grade 6 the IG received 400 hours of intensive ESL instruction over a 5-month period with the same teachers; (3) both groups reported having no extensive contact with English outside the classroom; (4) both groups attended the same secondary school, followed the same academic program with the same teachers and were also exposed to the same communicative-based ESL curricu-

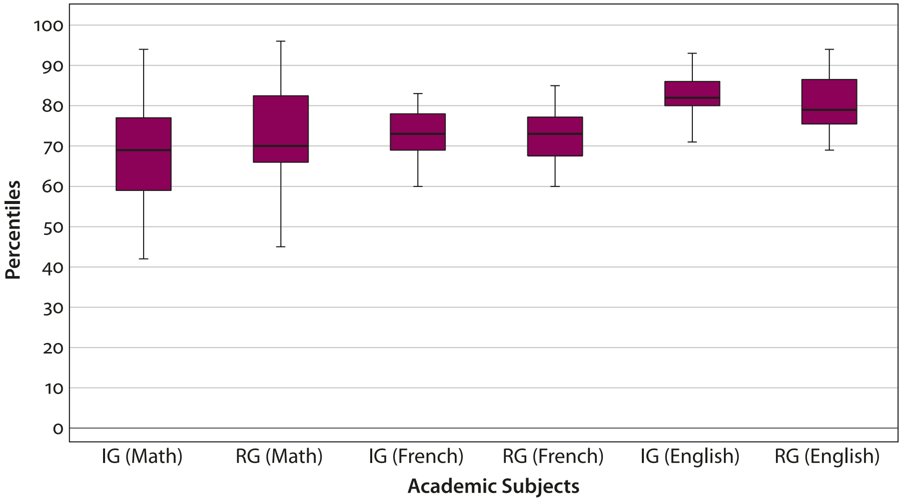


Figure 1. Academic performance (AP) in mathematics, French and English
Note. Maximum score = 100. RG (*n* = 39); IG (*n* = 42)

lum with the same teachers over a 4-year period; and (5) both groups were statistically similar in terms of academic performance and English proficiency at the time of data collection.

2.2 Procedure

Once we determined that both groups were statistically comparable in terms of academic performance and English proficiency, students completed a background questionnaire and a picture-cue narrative task. We administered the questionnaire to whole groups during a 30-minute period on the same day. We recorded individual students' narrative tasks over two days. Including instructions and clarifications, each recording session took approximately 10 minutes. Finally, we prepared speech samples from each recording for objective and perceived measures of oral production skills.

2.3 Background questionnaire (L2 contact)

We administered a student background questionnaire, based on French and O'Brien (2008) and Lightbown and Spada (1987), in French via Survey Monkey. The questionnaire elicited specific information about the nature and amount of English language/culture contact participants had had outside the classroom throughout primary and secondary school.

2.4 Speech elicitation task

To elicit speech samples, we used a picture narrative task (*The Suitcase Story)* (Derwing et al., 2004) that has produced reliable results in previous L2 studies examining different facets of speech production, including temporal fluency, perceived fluency, comprehensibility and accentedness (e.g. Derwing & Munro, 2013; Derwing et al., 2008; Derwing et al., 2009; Rossiter, Derwing, & Jones, 2008). In this task, participants saw an 8-picture storyline depicting a man and woman carrying identical green suitcases. While walking toward the corner of a busy city street, they accidentally run into one another and fall to the ground. After standing back up, they mistakenly retrieve the wrong suitcases and only discover the error when they arrive at their respective hotel rooms. Participants had one minute to familiarize themselves with the pictures and ask clarification questions prior to recounting their story. We digitally recorded the narratives in a quiet room for later analysis using Zoom H2 equipment.

2.5 Speech sample preparation

First, we converted the recordings to .wav files for analysis. From these recordings, we prepared 32-sec speech samples (i.e., the mean length of all narratives) from each participant ($N=81$), starting where the speaker began the narrative and excluding initial disfluencies (e.g., false starts, hesitations, etc.). We chose to examine the beginning of each narrative because we observed that speakers produced more idea units in the first third of the narrative. In doing so, the production content could be held relatively consistent across speakers when making comparisons. This approach also parallels previous oral production research, showing that speakers' L2 performance contains less variation during the beginning portions of a narrative task (e.g., Derwing, Thomson, & Munro, 2006).

The 81 speech samples (mean duration = 32 S) from the IG and RG were randomized, and we added four NS samples to verify that listeners used the entire Likert scale. We converted the finalized samples into a single sound file, which we then presented to listeners through a computer console using Sony MDR Headphones.

2.6 Raters

We recruited a total of five NSs of Canadian English (4 females; 1 male; age range 36–54; M = 36) with self-reported normal hearing to evaluate the French-speaking participants' fluency, comprehensibility, and accentedness, using 9-point Likert scales, which previous studies examining the tri-dimensional model of oral production have validated (e.g., Derwing, Rossiter, Munro & Thomson, 2004; Derwing et al., 2008, Derwing & Munro, 2013). All raters could be considered expert listeners, as each held a BA degree in second language teaching, and four held a graduate degree in Applied Linguistics, specializing in second language acquisition. All also had extensive teaching experience as ESL teachers in the public school system (experience range 6 – 17 years; M = 9.4) and had taught both in Quebec's regular (limited-exposure) and intensive ESL programs. The raters' education and professional profile, therefore, ensured that they had good familiarity with the language skills, and most importantly, the oral production skills of the groups under investigation in the present context.

2.7 Rating task

The five expert listeners rated the speech samples using separate 9-point scales for fluency (1 = *very fluent*, 9 = *not at all fluent*), comprehensibility (1 = *very easy to understand*, 9 = *very difficult to understand*) and accentedness (1 = *no accent*, 9 = *extremely heavy accent*). Following the work of Derwing and her colleagues (e.g., Derwing & Munro, 2013), we described each construct under evaluation to the raters. In particular, we instructed them not to consider language proficiency (i.e., use of grammar and vocabulary) when judging fluency and to base their judgements solely on the temporal aspects or overall fluidity of speech delivery (e.g., filled and silent pauses, self-repairs, speech rate, etc.). For comprehensibility, raters considered the cognitive effort required to understand a speech sample and then judged how easy or difficult it was to understand. Finally, for accentedness, raters judged how different speakers' accent was from that of standard Canadian English. We specifically asked listeners not to judge unidiomatic language use and/or grammatical and lexical inaccuracies but instead to focus their ratings entirely on the phonological aspects of productions (e.g., phonotactic properties, intonation, stress, etc.).

Before rating the stimulus set, raters received a copy of the 8-frame storyline, written definitions of all constructs and detailed instructions on how to use the scales. They then participated in a 15-minute training session during which they rated four practice items. The practice items were also 32-s speech samples of the narrative task and consisted of one NS and three L2 speakers (not part of the study) specifically chosen to reflect varying degrees of oral skill. This allowed listeners to experience firsthand a substantial range of oral proficiency (from L2 beginners to NSs). The ratings from this practice session ranged from 1–9 for all three constructs (fluency, comprehensibility, accentedness), demonstrating that listeners were indeed able to use the entire Likert scale to note differences in oral production ability prior to evaluating study samples. Immediately following the training session, listeners were equipped with Sony MDR Headphones and given a single booklet to record their written ratings. They then rated the same randomized set of speech samples ($N = 81$) together in a group session lasting 1.5 hours. During the session, listeners were given a mandatory 10-minute break every 30 minutes to reduce fatigue.

3. Results

3.1 Interrater reliability

To establish inter-rater reliability, we first determined that all raters had assigned a score of 1 to the four NS samples for each of the constructs under investigation (e.g., fluency, comprehensibility, accentedness). This indicated that they were consistent during the 1.5-hour rating session and, as observed in the training session, had also used the entire Likert scale. Since the raters were in total agreement about the NS samples, we removed these from the data set to avoid inflating interrater reliability estimates. We then computed a series of Cronbach's α across the five listeners, which revealed strong interrater reliability for ratings of fluency (0.94), comprehensibility (0.93) and accentedness (0.94). The magnitude of these estimates was consistent with those reported in previous research using perceptual measures of speech production in similar contexts (e.g., Derwing and Munro, 1997; Derwing et al., 2004; Derwing, Thomson, & Munro, 2006).

3.2 Fluency, comprehensibility and accentedness scores

Having established strong interrater reliability, we calculated the mean rating scores for fluency, comprehensibility and accentedness in both the IG and RG. We then compared these means in a series of independent *t*-tests (Bonferroni adjusted to require $p < .02$) and computed corresponding effect sizes (Cohen's *d*). The results (Table 2) revealed that NSs judged the IG to be significantly better than the RG in fluency ($t(79) = -.386$, $p < .02$, $d = .85$) and comprehensibility ($t(79) = -4.11$, $p < .02$, $d = .91$); there were also strong effect sizes for the observed group differences, particularly for comprehensibility (.91). However, the mean ratings for accentedness were virtually the same for IG and RG and did not differ significantly ($t(79) = -0.175$, $p = .86$, $d = .03$), pointing to no specific group advantage in terms of degrees of accentedness.

Table 2. Expert listeners' ratings of fluency, comprehensibility and accentedness: Intensive and regular groups

	IG		RG		*t (79)*	*p*	Cohen's *d*
	M	*SD*	*M*	*SD*			
Fluency	3.94	1.24	5.25	1.79	0.386	<.020	0.850
Comprehensibility	3.89	1.16	5.23	1.74	−4.110	<.020	0.910
Accentedness	5.70	0.91	5.74	1.09	−0.175	.860	0.030

Note. Intensive ($n=42$), Regular ($n=39$). Ratings on 9-point scale. *Accent: 1=no accent; Fluency: 1=very fluent; Comprehensibility: 1=very easy to understand.*

3.3 Utterance fluency measures

The NS judgements in Table 2 provide strong evidence of an advantage for the IG with respect to perceptual measures of fluency and comprehensibility, indicating that certain characteristics of this group's speech production were clearly different from that of the RG. As previous research has shown strong connections between temporal measures and fluency ratings (e.g., Derwing et al., 2004; Rossiter, 2009), we decided to examine the objective temporal aspects of both groups' speech samples in an attempt to gain additional insights into how the fluidity of speech delivery or "utterance fluency" (Segalowitz, 2010, 2016) may have affected raters' overall impressions of learners' fluency.

To obtain measures of utterance fluency, we first transcribed all 81 speech samples orthographically and then manually segmented these using Praat (Boersma & Weenink, 2012) to allow the visualization of speech waveforms. We then calculated speech segment durations using an automated script adapted from Préfontaine, Kormos, & Johnson (2016). This particular analysis produced a variety of different objective temporal measures related to utterance fluency for each speech sample; however, for the purpose of this study, we operationally defined utterance fluency as measures of (1) speech rate (number of syllables per second), (2) phonation time (percentage of time spent speaking) and (3) mean length of run (number of syllables in a run without filled or silent pauses). Previous studies have repeatedly shown that these particular measures provide a robust measure of L2 fluency in different contexts (Freed, 1995; Ginther, Slobadanka & Yang, 2010; Hilton, 2008; Kormos & Dénes, 2004; Mora & Valls-Ferrer, 2012; Segalowitz, 2010; Segalowitz, French, & Guay, 2017; Towell, Hawkins, & Bazergui, 1996).

To determine potential between-group differences on temporal measures, we conducted independent Bonferroni-adjusted *t*-tests (with the criterion for significance set to $p<.02$). We computed Cohen's *d* to assess effect sizes. Comparisons in

Table 3. Temporal measures of oral fluency

	IG		RG		*t (79)*	*p*	Cohen's *d*
	M	*SD*	*M*	*SD*			
Speech Rate	2.10	0.43	1.88	0.46	2.180	<.020	0.490
Mean Length of Run	4.52	1.66	3.61	1.22	3.180	<.020	0.680
Phonation Time (%)	59.10	10.91	51.41	14.95	2.030	<.020	0.590

Note. Intensive ($n = 42$), Regular ($n = 39$). Ratings on 9-point scale. Speech rate = number of syllables per second; Mean length of run = number of syllables in a run without filled or silent pauses; Phonation time = percentage of time spent speaking.

Table 2 show that the IG's mean speech rate, mean length of run and mean phonation time were all significantly greater than the RG ($t(79) = 2.18$, $p < .02$, $d = .49$; $t(79) = 3.18$, $p < .02$, $d = .68$; $t(79) = 2.03$, $p < .02$, $d = .59$, respectively). There were also moderate effect sizes for all three object temporal measures, suggesting further that, as was the case with perceived measures of fluency discussed above, IG's utterance fluency was also better than that of the RG, at least for the specific temporal measures examined in the present context.

Finally, we also computed Spearman correlations to examine the specific connections between different dimensions of oral production. We found that perceived fluency was strongly and significantly correlated with comprehensibility ($r(79) = .898$, $p < .01$). The correlation between fluency and accentedness was visibly lower ($r(79) = .491$, $p < .01$), but also similar in strength to the correlation between comprehensibility and accentedness ($r(79) = .510$, $p < .01$). We also computed correlations between the temporal measures of utterance fluency and perceived fluency. In this case, listeners' perceptions of fluency were significantly correlated with speech rate ($r(79) = .816$, $p < .01$), mean length of run ($r(79) = .655$, $p < .01$) and phonation time ($r(79) = .804$, $p < .01$), indicating a strong link between objective and perceived fluency measures.

4. Discussion

The main goal of this study was to examine long-term effects of intensive instruction on perceived measures of fluency, comprehensibility and accentedness in two groups of Grade 10 French-speaking students from the same high school. One group (IG) had received approximately 400 hours of intensive English instruction with the same teacher over a 5-month period in Grade 6 (age 11–12), whereas the other group (RG) had received regular drip-feed instruction (approximately 60 hrs) spread over the same school year. In high school, both groups returned to the same regimen of formal drip-feed instruction (approximately 100 hours per year) over a 4-year period. We accounted for the previously-reported confounding factors of academic ability and L2 proficiency in our study by comparing groups using school-reported scores for Math, French, and L2 oral proficiency. The groups were virtually identical across these measures. To achieve our main goal, we collected speech production data from both groups using a picture-cue narrative task. Expert listeners rated students' oral production for fluency, comprehensibility, and accentedness.

The overall results revealed that the expert listeners found the IG to be superior to the RG in terms of perceived fluency and comprehensibility; in fact, the effect sizes showed a strong advantage for the IG on these specific dimensions. Secondary analysis of temporal speech measures also revealed that the IG's utterance fluency (speech rate, mean length of run, phonation time) was significantly better than the RG, further validating listeners' perceptions that the IG demonstrated superior fluency skills. There was, however, no difference in listeners' perceptions of accentedness, which suggests that both groups had similar pronunciation skills.

In the following section, we discuss the major findings and propose explanations. We argue that the IG's advantage for perceived fluency, utterance fluency, and comprehensibility appear to be the result of a critical threshold of skill proceduralization that was previously attained in their intensive program. We further argue that the absence of between-group differences in terms of accentedness is largely due to the lack of explicit teaching of pronunciation in the classroom.

4.1 Fluency

The findings from both perceived and utterance fluency showed that the IG learners were considerably more fluent than students who had only received drip-feed instruction in elementary school. In fact, 4 years after their intensive program, when compared to their age peers, IG learners spoke faster, produced longer, lexically-dense utterances, and could speak for longer intervals without pausing. NSs also reported that their speech flowed much better than their peers. Not only do these findings provide additional evidence that intensive instruction promotes the development of L2 oral fluency (Freed et al., 2004; Lightbown & Spada, 1994; Serrano, Llanes, & Tragant, 2016) but they also provide new evidence that resulting fluency benefits may persist over time (up to 4 years), even after taking into account potentially mediating factors such as academic ability and L2 proficiency.

The presence of a long-term fluency advantage in the present context is also quite consistent with that reported in Lightbown and Spada (1991). Using measures of oral production targeting the notion of "talkativeness" (referenced as speech volume and number of extended turns in an interview), they found that, 5 years after their intensive program, intensive learners were more talkative and could hold the floor longer in conversations than their age peers who had received no intensive instruction. In the present context, although the oral production task was different, specific measures of utterance fluency (phonation time and mean length of run) also showed that the IG spent more time "talking" than "pausing" when producing their picture-cue narratives. What emerges then from both these research contexts is that IG learners maintained a considerable fluency advantage over their peers, suggesting that gains in L2 fluency associated with previous intensive learning appear to have continued even after returning to several years of drip-feed instruction.

An important question that follows, then, is why intensive instruction may have led to long-term benefits for learners' fluency. One obvious explanation is the role of practice. Intensive programs, through concentrated instruction over a short period of time, provide numerous opportunities for intensive and frequent L2 practice. This, in turn, creates a rich learning context for building procedural knowledge (DeKeyser, 2007, 2015). It may, therefore, be that the degree of intensity of language use and practice in this context is of particular benefit to the development of utterance fluency because its underlying temporal elements (e.g., speech rate, mean length of run, phonation time) seem to be quite sensitive to effects of proceduralization (Goldman Eisler, 1968). As such, it could be in the present context that learners' utterance fluency had reached a critical threshold of proceduralization during their intensive experience, which made it far more robust to attrition over time (Huensch et al., 2019).

4.2 Comprehensibility

In this study, learners' degree of comprehensibility was a central distinguishing factor between the two groups, providing new evidence that NSs of English actually found IG learners to be much easier to understand than their peers. Moreover, as shown in previous research (e.g. Crowther, Trofimovich, Isaacs, & Saito, 2015; Derwing et al., 2008; Derwing et al., 2004; Isaacs & Trofimovich, 2012), it is likely that perceptions of comprehensibility were tied to learners' fluency. In fact, perceived measures of fluency in the present context were strongly correlated with listeners' ratings for comprehensibility, suggesting that IG learners were easier to understand most likely because they spoke faster, paused less frequently and used longer linguistically-dense utterances. These findings also suggest that both fluency and comprehensibility were interrelated and thus developed together over time. Consequently, it is likely that IG learners were also able to maintain a long-term comprehensibility advantage because of the strong connection it shared with their existing fluency skills.

However, although there is evidence of a strong long-term connection between learners' fluency and comprehensibility in the present context, it is interesting to note that previous studies (e.g., Isaacs & Trofimovich, 2012; Munro & Derwing, 1995b) have consistently shown that grammatical and lexical errors can also influence the degree of comprehensibility. It is therefore possible that the IG's speech may have been more accurate, both grammatically and lexically, which, in addition to fluency skills, could have at least partially contributed to their long-term advantage in comprehensibility. A lexical and grammatical comparison of the speech produced by IG learners and their peers was outside the scope of the present study; however, additional longitudinal research in intensive learning settings would clearly benefit from examining how degrees of accuracy might moderate the long-term developmental relationship between fluency and comprehensibility.

4.3 Accentedness

The NSs of English in this context rated both the IG learners and their age peers virtually identically in terms of their L2 accent; in fact, unlike for fluency and comprehensibility, there was no group advantage whatsoever for the degree of accentedness, revealing a partial independence among speech dimensions. Similar to other developmental studies that have reported little accent development over time (e.g., Derwing & Munro, 2013), both groups of learners in the present study focused primarily on meaning-focused oral communication. Classroom instruction, while strongly encouraging communication, provided little explicit attention to form, especially to phonemic and prosodic features that do not necessarily impede comprehensibility because of their low functional load (Derwing & Munro, 2015). It is therefore not surprising that both groups presented similar degrees of accentedness. Given this finding, if learners wish to reduce accentedness, they may need to focus on these low functional load features. However, as we observed in the present context, because accentedness did not necessarily have a significant impact on the development of comprehensibility, it may indeed be far more beneficial to focus classroom instruction on developing pronunciation features that promote improved comprehensibility (e.g., word stress). Additional research would shed further light on the long-term effects of this type of awareness-raising instruction on comprehensibility.

Nevertheless, despite both groups having similar strongly accented speech (i.e., as indicated by an average rating score of almost 6 on the 9-point scale), listeners still reported that IG learners spoke more smoothly and were easier to understand than their peers, suggesting that degree of accentedness had little influence on listeners' perception of fluency and comprehensibility. Derwing and her colleagues have repeatedly found that perceptions of comprehensibility are more closely linked to fluency than to those of accentedness (Derwing & Munro, 1997; Munro & Derwing, 1999). Our findings of a strong link between perceptions of fluency and comprehensibility and an even smaller connection between comprehensibility and accentedness are therefore clearly consistent with these accounts and further underscore the positive impact of intensive instruction on the development of learners' perceived fluency and comprehensibility.

5. Conclusion

As we noted at the onset of this study, very little is known about the long-term benefits of intensive instruction. We examined whether high school students who had received intensive instruction in grade school showed any advantage on oral production skills (fluency, comprehensibility, accentedness) when compared to their age peers who had only been exposed to drip-feed L2 instruction. The findings clearly showed that IG learners held a strong advantage over their RG peers for both fluency and comprehensibility, even after accounting for the effects of academic ability and overall L2 proficiency. In fact, IG learners actually spoke faster and more smoothly, produced longer pause-free sentences, and were much easier to understand than RG learners. Both groups of learners, however, exhibited similar degrees of accentedness.

Overall, then, our findings suggest that, through intensive and frequent L2 practice, IG learners' fluency skill, in particular, appears to reach a critical threshold of proceduralization during the intensive program, which in turn produces long-term benefits for both fluency and comprehensibility. Learners' accentedness, on the other hand, undergoes virtually no development over time, resulting most likely from a lack of systematic pedagogical focus on the phonemic and prosodic features of pronunciation. Our findings also provide confirmation that Derwing and colleagues' tridimensional model of oral production can be used to distinguish useful perceptual nuances in young adolescents' L2 oral skills. Finally, through the use of expert speaker perceptions of comprehensibility, our findings provide new evidence about the extent to which English NSs are able to understand young ESL learners, which furthers our understanding of how IG learners' oral skills are actually viewed outside the educational context.

This study, however, is not without limitations. First, although we matched learners on academic ability and L2 proficiency in Grade 10, we were not able to assess all three dimensions of IG learners' oral production at the end of their intensive program in Grade 6, making it impossible to determine whether fluctuations in their fluency, comprehensibility and accentedness may have occurred over time between Grade 6 and Grade 10. This highlights the need to carry out longitudinal research with IG learners that not only controls for confounding factors but also assesses the development of oral skills at yearly intervals, which would provide further understanding of how such skills are maintained or lost over time. Moreover, our findings were based on speech samples elicited from a single (monologic) narrative task. However, oral performance, and in particular fluency, can vary considerably across task types (Crowther, 2020; Tavokoli, 2016); additional research would therefore help to determine whether IG learners' performance on different oral tasks (e.g. monologic versus dialogic) also shows long-term benefits similar to those in the present study. Finally, our findings are based on learners who had very little access to the L2 outside the classroom, which created a rather homogenous learning context. It would, therefore, be important to examine whether IG learners' advantage for fluency is also maintained over RG peers in an L2 learning context where learners have continual opportunities in their community to engage in meaningful L2 practice.

Nevertheless, despite certain limitations, this study provides convincing evidence that an intensive learning experience produces long-term benefits for L2 learners' fluency and comprehensibility, which adds to a growing body of research on the effects of intensive instruction. Given the central findings of the current research, it may be of particular benefit to have schools create follow-up programs designed specifically to help IG students maintain their fluency and related skills as well as develop L2 knowledge in other areas. It may also be necessary for language educators to strongly consider the possibility of replacing more traditional drip-feed curricula with different models of intensive L2 instruction as a means to boost successful L2 learning outcomes for all learners. Answers to these questions and those raised previously will not only be of interest to researchers but will also be quite pertinent for school administrators and language planners as they look for new ways to optimize L2 learners' fluency, comprehensibility and accentedness over time.

References

Boersma, P., & Weenink, D. (2012). Praat: Doing phonetics by computer, version 5.3.53. http://www.praat.org/

Collins, L., Lightbown, P.M., Halter, R.H., & Spada, N.M. (1999). Time and the distribution of time in L2 instruction. *TESOL Quarterly*, 33(4), 655–680. https://doi.org/10.2307/3587881

Collins, L., & Muñoz, C. (2016). The foreign language classroom: Current perspectives and future considerations. *The Modern Language Journal*, 100(S1), 133–147. https://doi.org/10.1111/modl.12305

Collins, L., & White, J. (2011). An intensive look at intensity and language learning. *TESOL Quarterly*, 45(1), 107–133. https://doi.org/10.5054/tq.2011.240858

Collins, L., & White, J. (2012). Closing the gap: Intensity and proficiency. In C. Munoz (Ed.), *Intensive exposure experiences in second language learning*. Bristol: Multilingual Matters. https://doi.org/10.21832/9781847698063-006

Crowther, D. (2020). Rating L2 speaker comprehensibility on monologic vs. interactive tasks: What is the effect of speaking task type? *Journal of Second Language Pronunciation*, 6(1), 96–121. https://doi.org/10.1075/jslp.19019.cro

Crowther, D., Trofimovich, P., Isaacs, T., & Saito, K. (2015). Does a speaking task affect second language comprehensibility? *The Modern Language Journal*, 99(1), 80–95. https://doi.org/10.1111/modl.12185

DeKeyser, R. (2007). Study abroad as foreign language practice. In R. DeKeyser (Ed.), *Practice in a second language: Perspectives from applied linguistics and cognitive psychology*. New York, NY: Cambridge University Press. https://doi.org/10.1017/CBO9780511667275.012

DeKeyser, R. (2015). Skill acquisition theory. In B. VanPatten & J.N. William (Eds.), *Theories in second language acquisition. An introduction* (pp. 94–112). London: Routledge.

Derwing, T.M., & Munro, M.J. (1997). Accent, intelligibility and comprehensibility. *Studies in Second Language Acquisition*, 19(01), 1–16. https://doi.org/10.1017/S0272263197001010

Derwing, T.M., & Munro, M.J. (2015). *Pronunciation fundamentals. Evidenced-based perspectives for L2 teaching and research*. Philadelphia, PA: John Benjamins Publishing Company. https://doi.org/10.1075/lllt.42

Derwing, T.M., & Munro, M.J. (2013). The development of L2 oral language skills in two L1 groups: A 7-year study. *Language Learning*, 63(2), 163–185. https://doi.org/10.1111/lang.12000

Derwing, T.M., Munro, M.J., & Thomson, R.I. (2008). A longitudinal study of ESL learners' fluency and comprehensibility development. *Applied Linguistics*, 29(3), 359–380. https://doi.org/10.1093/applin/amm041

Derwing, T.M., Munro, M.J., Thomson, R.I., & Rossiter, M.J. (2009). The relationship between L1 fluency and L2 fluency development. *Studies in Second Language Acquisition*, 31(04), 533–557. https://doi.org/10.1017/S0272263109990015

Derwing, T.M., Rossiter, M.J., Munro, J.M., & Thomson, R.I. (2004). Second language fluency: Judgments on different tasks. *Language Learning*, 54, 655–679. https://doi.org/10.1111/j.1467-9922.2004.00282.x

Derwing, T.M., Thomson, R.I., & Munro, J.M. (2006). English pronunciation and fluency development in Mandarin and Slavic speakers. *System*, 34, 183–193. https://doi.org/10.1016/j.system.2006.01.005

Freed, B.F. (1995). *Language learning and study abroad. Second language acquisition in a study abroad context*, 123–148. https://doi.org/10.1075/sibil.9.09fre

Freed, B.F., Segalowitz, N., & Dewey, D.P. (2004). Comparing regular classroom, study abroad, and intensive domestic immersion programs. *Studies in Second Language Acquisition*, 26, 275–301. https://doi.org/10.1017/S0272263104262064

French, L.M., & O'Brien, I. (2008). Phonological memory and children's second language grammar learning. *Applied Psycholinguistics*, 29(3), 463–487. https://doi.org/10.1017/S0142716408080211

Ginther, A., Slobadanka, D., et Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing*, 27(3), 379–399. https://doi.org/10.1177/0265532210364407

Goldman Eisler, F. (1968). *Psycholinguistics: Experiments in spontaneous speech*. University of Michigan: Academic P.

Hilton, H. (2008). The link between vocabulary knowledge and spoken L2 fluency. *Language Learning Journal*, 36(2), 153–166. https://doi.org/10.1080/09571730802389983

Huensch, A., & Tracy-Ventura, N. (2017a). Understanding second language fluency behavior: The effects of individual differences in first language fluency, cross-linguistic differences, and proficiency over time. *Applied Psycholinguistics*, 38. https://doi.org/10.1017/S0142716416000424

Huensch, A., Tracy-Ventura, N., Bridges, J., & Cuesta Medina, J.A. (2019). Variables affecting the maintenance of L2 proficiency and fluency four years post-study abroad. *Study Abroad Research in Second Language Acquisition and International Education*, 4(1), 96–125. https://doi.org/10.1075/sar.17015.hue

Huensch, A., & Tracy–Ventura, N. (2017b). L2 utterance fluency development before, during, and after residence abroad: A multidimensional investigation. *The Modern Language Journal*, 101(2), 275–293. https://doi.org/10.1111/modl.12395

Institut de la statistique du Québec. (2011). Population selon la langue maternelle, régions administratives et ensemble du Québec, 2011. Retrieved from http://www.stat.gouv.qc.ca/statistiques/recensement

Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition* 34, 475–505. https://doi.org/10.1017/S0272263112000150

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, 32, 145–164. https://doi.org/10.1016/j.system.2004.01.001

Lightbown, P.M. (2014). Making the minutes count in L2 teaching. *Language Awareness*, 1–2(23), 3–23. https://doi.org/10.1080/09658416.2013.863903

Lightbown, P.M., & Spada, N.M. (1987). *Learning English in intensive programs in Quebec schools (1986–1987): Report of the first year of research*. Montreal: Concordia University.

Lightbown, P.M., & Spada, N.M. (1989). *A secondary V follow-up study of learners from primary-level intensive ESL programs (research report)*. Montréal: Concordia University.

Lightbown, P.M., & Spada, N.M. (1991). Étude des effets à long terme de l'apprentissage intensif de l'anglais, langue seconde, au primaire. *La revue canadienne des langues vivantes*(48), 90–117. https://doi.org/10.3138/cmlr.48.1.90

Lightbown, P.M., & Spada, N.M. (1994). An innovative program for primary ESL students in Quebec. *TESOL Quarterly*, 28(3), 563–579. https://doi.org/10.2307/3587308

Llanes, A. (2012). The short- and long-term effects of a short study abroad experience: The case of children. *System*, 40(2), 179–190. https://doi.org/10.1016/j.system.2012.05.003

Mora, J.C., & Valls-Ferrer, M. (2012). Oral fluency, accuracy, and complexity in formal instruction and study abroad learning contexts. *TESOL Quarterly*, 1–32. https://doi.org/10.1002/tesq.34

Muñoz, C. (2012). *Intensive exposure experiences in second language learning*. Bristol: Multilingual Matters. https://doi.org/10.21832/9781847698063

Munro, M. J., & Derwing, T. M. (1995a). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, 49, 285–310. https://doi.org/10.1111/0023-8333.49.s1.8

Munro, M. J., & Derwing, T. M. (1995b). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech*, 38(3), 289–306. https://doi.org/10.1177/002383099503800305

Munro, M. J., & Derwing, T. M. (1999). Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language Learners. *Language Learning*, 49(s1), 285–310. https://doi.org/10.1111/0023-8333.49.s1.8

Préfontaine, Y., Kormos, J., & Johnson, D. E. (2016). How do utterance measures predict raters' perceptions of fluency in French as a second language? *Language Testing*, 33(1), 53–73. https://doi.org/10.1177/0265532215579530

Regan, V. (2005). From speech community back to classroom: What variation analysis can tell us about the role of context in the acquisition of French as a foreign language. *Focus on French as a foreign language: Multidisciplinary perspectives*, 191–209. https://doi.org/10.21832/9781853597688-010

Rossiter, M. J. (2009). Perceptions of L2 fluency by native and non-native speakers of English. *The Canadian Modern Language Review*, 65(3), 395–412. https://doi.org/10.3138/cmlr.65.3.395

Rossiter, M. J., Derwing, T. M., & Jones, V. M. (2008). Is a picture worth a thousand words? *TESOL Quarterly*, 42(2), 325–329. https://doi.org/10.1002/j.1545-7249.2008.tb00127.x

Sato, T., & McNamara, T. (2018). What counts in second language oral communication ability? The perspective of linguistic laypersons. *Applied Linguistics*, 40(6), 894–916. https://doi.org/10.1093/applin/amy032

Segalowitz, N. (2010). *The cognitive bases of second language fluency*. New York: Routledge. https://doi.org/10.4324/9780203851357

Segalowitz, N. (2016). Second language fluency and its underlying cognitive and social determinants. *IRAL: International Review of Applied Linguistics in Language Teaching*, 54(2), 79–95. https://doi.org/10.1515/iral-2016-9991

Segalowitz, N., French, L., & Guay, J.-D. (2017). What features best characterize adult second language utterance fluency and what do they reveal about fluency gains in short-term immersion? *Canadian Journal of Applied Linguistics*, 20, 90–116. https://doi.org/10.7202/1050813ar

Serrano, R. (2012). Is intensive learning effective? Reflecting on the results from cognitive psychology and the second language acquisition literature. In C. Muñoz (Ed.), *Intensive Exposure Experiences in second language learning* (pp. 3–24). Bristol: Multilingual Matters. https://doi.org/10.21832/9781847698063-004

Serrano, R., Llanes, A., & Tragant, E. (2016). Examining L2 development in two short-term intensive programs for teenagers: Study abroad vs. "at home". *System*, 57, 43–54. https://doi.org/10.1016/j.system.2016.01.003

Serrano, R., & Muñoz, C. (2007). Same hours, different time distribution: Any difference in EFL? *System*, 35, 305–321. https://doi.org/10.1016/j.system.2007.02.001

Tavakoli, P. (2016). Fluency in monologic and dialogic task performance: Challenges in defining and measuring L2 fluency. *IRAL: International Review of Applied Linguistics in Language Teaching*, 54(2), 133–150. https://doi.org/10.1515/iral-2016-9994

Tomczak, M., & Tomczak, E. (2014). The need to report effect size estimates revisited. An overview of some recommended measures of effect size. *Trends in Sport Sciences*, 1(21), 19–25.

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced of French. . *Applied Linguistics*, 17(1), 84–119. https://doi.org/10.1093/applin/17.1.84

White, J. L., & Turner, C. E. (2005). Comparing children's oral ability in two ESL programs. *Canadian Modern Language Review*, 61(4), 491–517. https://doi.org/10.3138/cmlr.61.4.491

Reactions to second language speech[*]

Influences of discrete speech characteristics, rater experience, and speaker first language background

Talia Isaacs and Ron I. Thomson
University College London | Brock University

This study investigates how Mandarin and Slavic language speakers' comprehensibility, accentedness, and fluency ratings, as assigned by experienced teacher-raters and novice raters, align with discrete linguistic measures, and raters' accounts of influences on their scoring. In addition to examining mean ratings in relation to rater experience and speaker first language background, we correlated ratings with segmental, prosodic, and temporal measures. Introspective reports were segmented, coded, enumerated, and submitted to loglinear analysis to elucidate influences on ratings. Results showed that ratings were strongly correlated with prosodic goodness and moderately correlated with segmental errors, implying the importance of both segmentals and prosody in L2 speech ratings. Experienced teacher-raters provided lengthier reports than novice raters, producing more comments for all coded categories where an error was identified except for pausing (a disfluency marker). This may be because novice raters observed little else about the speech or struggled to pinpoint or articulate other features.

Keywords: accent, comprehensibility, English as a second language, fluency, pronunciation assessment, raters, rating scales, speech perception

1. Introduction

A growing body of second language (L2) pronunciation research examining global perceptual constructs (e.g., comprehensibility, accentedness, fluency) in

[*] An earlier version of this article was published as part of a special issue of the *Journal of Second Language Pronunciation* 6(3). https://doi.org/10.1075/jslp.20018.isa

https://doi.org/10.1075/bct.121.07isa

relation to discrete linguistic measures (e.g., segmental accuracy, temporal measures) has exerted a sustained influence on L2 speaking assessment research over the past decade (Isaacs & Harding, 2017). If we accept the view that both speakers and listeners play a role in successfully exchanging oral messages (Schiavetti, 1992) and share communicative responsibility (Rajadurai, 2007), a few points logically follow. This includes needing to better understand what features of L2 speech are salient to different types of listeners. We also need to examine whether listeners' beliefs about which linguistic features inform their assessments match what is actually present in learner speech.

In traditional L2 pronunciation research, ratings of global perceptual constructs are often measured using 9-point numerical scales, with brief, relativistic descriptors anchoring the scales on each end (e.g., no accent/extremely strong accent; Derwing & Munro, 1997). These scales have the advantage of being user-friendly, jargon-free, and accessible to raters who may lack specialist knowledge of pronunciation. Further, ratings obtained using these Likert-type scales consistently yield high interrater reliability across studies, even without listener training (Munro, 2018). However, such scales provide raters with little guidance on how to interpret score levels. Even if there is exact rater agreement on a score assigned to an L2 speaking performance, it does not *necessarily* follow that raters arrived at the same score for the same reasons or interpreted the constructs in the same way (Douglas, 1994). Indeed, a fundamental principle in psychometrics is that reliability is a prerequisite for construct validity but is an insufficient condition for it (Bannigan & Watson, 2009). Therefore, it is important to establish what lies beneath listeners' impressionistic judgments and scoring decisions.

Variability is integral to the rating process, with ratings of speech involving both L2 learners and raters who vary on many characteristics (e.g., cognitive, attitudinal). Raters interact with the speech elicitation task and scoring system in different ways to generate a score (Upshur & Turner, 1999). If numerous deviations from native patterns were to co-occur in a speech sample, raters may tune into different constellations of deviations (Munro, 2018). They then need to filter their impressions through the artifact of a scoring system, with descriptors necessarily underrepresenting the complexity of performances (Lumley, 2005). Variability in L2 learner performance on the trait being measured is desirable, so that learners' ability levels can be differentiated and reflected in the scoring. The criteria that raters use to assign meaning to scale levels are important to investigate in research contexts, where, in contrast to many high-stakes assessment settings that use extended scale descriptors, raters receive scant guidance from rating scales and little rater training. Hence, they need to arrive at their own understanding of what the scale levels mean in terms of performance features during real-time scoring. To date, few L2 pronunciation studies have used introspective methods to probe listeners' accounts of influences on their scoring decisions. Derwing and Munro (2009) elicited listeners' written reports about preferences for L2 recorded voices, which had been pre-rated at different L2 comprehensibility and accentedness levels. Other researchers have used introspective reports to extend quantitative findings about the relationship between discrete linguistic measures and global L2 speech ratings (e.g., Foote & Trofimovich, 2018; Isaacs & Trofimovich, 2012).

The current study contributes to this emerging body of research, combining raters' verbalizations with other sources of evidence to illuminate their responses to L2 speech. More specifically, we analyze experienced teacher-raters' accounts compared to those of novice raters (undergraduate students) and how their ratings align with linguistic measures derived from the L2 speech samples. Eliciting ratings from experienced teacher-raters and novice listeners in settings where English is used as a lingua franca is ecologically valid due to likely interactions involving L2 speakers inside and/or outside of the classroom (Rose & Galloway, 2019), although only teachers would likely formally assess their speech.

The variability associated with rater experience is not viewed as a threat to validity in this study (see Isaacs & Thomson, 2013, for a discussion of the rater experience construct in L2 pronunciation research). Rather, it is regarded as a rich source of information that allows reflection on our understanding of global constructs often examined in pronunciation research (Chalhoub-Deville, 1995). Listeners are by far the best resource for better understanding such constructs, which, by definition, relate to listener perceptions of L2 speech. Thus, examining listeners' interpretations of the focal constructs, listening and rating processes and strategies, and how their perceptions align with linguistic characteristics of spoken productions (e.g., word choice, grammar) is essential for better understanding the L2 abilities we are attempting to measure.

2. The current study

This study brings together insights from two disciplines: language testing research on systematic sources of variance in human scoring, and L2 pronunciation research on the linguistic properties underlying global perceptual constructs. The goal is to examine the linguistic variables that underlie comprehensibility, accentedness, and fluency ratings. We examine how listeners' ratings align with both discrete L2 speech measures (e.g., segmental error counts, speaking rate), and listener reports of linguistic features that they attend to, grouped by listener experience and speaker first language (L1) background variables. These aims are distilled into the following research questions:

1. Which discrete L2 pronunciation and fluency measures are most related to listeners' global ratings of comprehensibility, accentedness, and fluency?
   - Does listener experience play a role?
   - Do learners' L1 backgrounds influence the listener?
2. How do listeners' perceptions of the linguistic influences on their judgments relate to these global L2 speech ratings?
   - Does listener experience play a role?
   - Do learners' L1 backgrounds influence the listener?

3. Method

3.1 Research design

In holistic rating, raters condense their impressions of a complex L2 performance into a single rating. Previous research has established that even highly trained raters may draw on different criteria to make scoring decisions, which may or may not be reflected in the scale descriptors (Lumley, 2005). Multiple sources of evidence were needed to elucidate this research problem. Therefore, a concurrent mixed methods design was used (Creswell & Plano Clark, 2017). To address the first research question, experienced teacher-raters' versus novice raters' global pronunciation and fluency ratings of L2 Mandarin and Slavic language speakers' utterances were statistically examined in relation to segmental, prosodic, and temporal measures. For research question two, an inductive coding scheme was generated from raters' introspective reports. The coded comments were then quantified and counts of coded categories for experienced teacher-raters versus novice listeners and Mandarin versus Slavic language speakers were obtained. The highest frequency codes were then subjected to quantitative analysis to test for between-group differences.

3.2 L2 speakers

Speech samples were elicited from 38 adult newcomers to Canada (27 females, 11 males, $M_{\text{age}} = 39.4$ years; 29–52). Half were L1 Mandarin speakers, who reported first exposure to English at a mean age of 14.3 years (7.0) and had resided in Canada for 16.7 months on average (11.9). The other half were L1 Slavic speakers (13 Russian, 3 Serbo-Croatian, 2 Ukrainian, 1 Polish), whose first reported English exposure was at a mean age of 16.2 years (11.8), with 15.6 months' Canadian residency on average (10.7). All were assessed at beginner English levels on the Canadian Language Benchmarks (CLB levels 1–4 of the instrument; Pawlikowska-Smith, 2000) and were enrolled in the government-funded Language Instruction for Newcomers to Canada (LINC) program at the time of the study. Mandarin and Slavic language speakers were matched for proficiency level based on the English as a Second Language (ESL) class in which they were registered. Placement decisions had been based on both CLB level and results from an in-house English proficiency test, which assessed L2 grammatical and lexical knowledge, literacy skills, and aural/oral performance.

Table 1 shows Mandarin and Slavic language speakers' self-reported L2 English exposure and estimated proficiency levels, obtained from questionnaire items administered at the beginning of data collection. Mandarin learners estimated speaking and listening to English outside class a greater proportion of the time than did Slavic language speakers but perceived having extended conversations with L1 English speakers less often and assessed their overall proficiency at a lower level. However, none of these self-report measures were statistically significant, $t(36) = |.19–1.78|$ $p > .05$, suggesting that the L1 groups were matched on language-related variables.

Table 1. Mandarin and Slavic language speakers' reported English language exposure and proficiency

Self-report measures	**L1 Mandarin**		**L1 Slavic**	
	M	*SD*	*M*	*SD*
Percent of time speaking English outside of class	35.8	25.5	34.2	26.5
Percent of time listening to English language media	80.5	24.6	71.6	33.5
Number of extended conversations with native English speakers per week[a]	1.8	2.4	2.5	3.0
English listening/speaking proficiency[b]	3.9	1.5	4.7	1.3
English reading/writing proficiency[b]	4.9	1.8	5.2	1.3

Notes.
[a] An extended conversation was defined as ≥10 min
[b] Measured on a 9-point Likert-type scale (*1 = extremely poor, 9 = extremely proficient*).

3.3 Speech elicitation and data preparation

Speech samples were audio recorded on several speaking tasks in a quiet room using a Marantz PMD661 SD recorder (duration: ≤ 40 mins). This article will report on performance on one task, an eight-frame picture narrative often used to elicit adult ESL learners' extemporaneous speech samples in L2 pronunciation and fluency research (Derwing & Munro, 2013). The essential plot elements were the collision of a man and a woman carrying similar suitcases on the street, their retrieval of the wrong suitcase and eventual discovery that they had accidently exchanged suitcases. The speakers were given a minute to look over the visual prompt before describing the picture sequence. After normalizing the speech samples for peak amplitude and removing any disfluencies that had preceded the storytelling (e.g., false starts, hesitations), the first 20 seconds of each narrative were excised from the recordings and randomized in preparation for rating ($M_{\text{duration}} = 27.1$ s; $SD = 2.3$). The speech sample of a male native English speaker was included about two thirds of the way through the set of recordings for all randomizations to verify that listeners' ratings corresponded to the correct speech sample in the printed response sheet. Once this was established, the native speaker's ratings were excluded from subsequent analyses.

3.4 Raters

Forty native English speakers, who reported having normal hearing, participated as raters. Half were experienced ESL teachers (14 females, 6 males; $M_{\text{experience}} = 9.7$ years; $SD = 5.1$), who either held or were pursing graduate degrees in applied linguistics from a Canadian English-medium university. These experienced teacher-raters reported teaching ESL for 13.9 hrs/week on average before commencing their studies ($SD = 8.47$). However, they varied in their teacher training, with 13 having taken a pronunciation course for teachers, 16 an L2 assessment course, and two with no training in these areas. The remaining 20 raters (15 females, 5 males), henceforth referred to as novice raters, were pursuing graduate degrees in nonlinguistic disciplines (e.g., political science, law, epidemiology) and uniformly had no assessment training.

The raters indicated their age range from a list in a background questionnaire due to some raters' sensitivity about age reporting during piloting. The experienced teacher-raters were the older demographic, with two raters in their 20s, 10 in their 30s, five in their 40s, and three in the 50 years or over age category. In contrast, 15 novice raters were in their 20s and only five were over 30. As a precondition for participating, only raters who reported never having learned Chinese or Russian (the most common Slavic L1 in the study) and who did not have notable exposure to members from either language community (e.g., through family relations, extended travel) could take part.

At the beginning of data collection, recruited raters were asked about their L1 accent familiarity in a background questionnaire (*1 = extremely unfamiliar, 9 = extremely familiar*). They reported significantly greater familiarity with Mandarin speakers' English ($M = 4.38$, $SD = 2.52$) than that of Russian speakers

(M = 3.28, SD = 2.21), $t(39)$ = 3.65; p = .001, with significant effects retained when raters were broken down into experienced teacher, $t(19)$ = 2.44; p = .025, and novice groups, $t(19)$ = 2.89; p = .009. Table 2 shows that experienced teacher-raters reported interacting significantly more with L2 speakers as a proportion of their total time than did novice raters, $t(38)$ = 3.02, $p < .005$. This is unsurprising, since teaching time was subsumed in experienced teacher-raters' estimates but was, by definition, absent from novice raters' estimates. Experienced teacher-raters also reported significantly greater exposure than novice raters to the English speech of both Mandarin learners, $t(38)$ = 3.15, $p < .001$, and Slavic language speakers, $t(38)$ = 2.20, $p < .002$.

Table 2. Experienced teacher-raters' and novice raters' self-reported mean interactions with L2 speakers and exposure to the L2 English of Mandarin and Slavic speakers

Self-report measures	**Experienced**		**Novice**	
	M	*SD*	*M*	*SD*
Percentage of time interacting with L2 speakers	39.0	16.83	22.5	17.73
Exposure to Mandarin-accented speech[a]	5.60	2.39	3.15	2.03
Exposure to Slavic-accented speech[a]	4.35	2.43	2.20	1.28

a Measured on a 9-point Likert-type scale (*1 = extremely familiar*, *9 = extremely unfamiliar*).

3.5 Rating sessions

The rating sessions were conducted individually in a quiet office, with a short break to mitigate rater fatigue (duration: ≤ 2 hrs). After hearing each speech sample, raters recorded scores on separate numerical scales for comprehensibility (very hard/very easy to understand), accentedness (heavily accented/not accented at all), and fluency (very disfluent/very fluent), with descriptors at scale anchors. As part of a larger study examining rating scale length (Isaacs & Thomson, 2013), half of each rater group was arbitrarily assigned to either a 5-point or 9-point rating scale length condition. In order to establish a baseline understanding about the constructs they were rating, we provided raters with explicit definitional guidance. Comprehensibility was defined as how easy the L2 speech is to understand (Derwing & Munro, 1997); accentedness denoted how different the speech sounds from that of a native speaker of North American English (Isaacs & Thomson, 2013); and fluency referred to the smoothness and rapidness of the oral delivery, corresponding to Lennon's (1990) narrow sense of the term and reflecting temporal phenomena (e.g., speech rate, hesitations). After familiarizing raters with the speaking prompt and rating procedures, they received general feedback on their ratings of four practice items (2 native English speakers, 1 Mandarin speaker, 1 Slavic language speaker) based on comparisons with mean scores that had previously been assigned by an independent group of raters in Derwing, Thomson, and Munro (2006). Specifically, they were told whether their ratings were considerably harsher, considerably more lenient, or roughly the same compared to mean scores assigned by the previous group of raters. In all cases, the researcher highlighted that there were no right or wrong answers and raters were not directed to adjust their scoring as a result.

Introspective reports were elicited for the linguistic factors that experienced teacher-raters and novice raters reportedly attended to when listening to and rating the speech (Gass & Mackey, 2000). Half of the raters in each rater group completed verbal protocols during their first listening. Procedurally, this involved the researcher pausing immediately after each recording so raters could articulate their thoughts while completing their ratings or reflecting on their scoring. If a halting silence occurred, the researcher prompted raters to continue verbalizing their thoughts with the probe, "what are you thinking?" However, raters were the ultimate arbiters of the amount of commentary they delivered, indicating when they were ready to proceed to the next recording using verbal or nonverbal signals (M_{duration} of listening, rating, and verbal protocols = 39 min and 34 min for experienced teacher-raters and novice raters respectively, range: 25–57 min). Because the additional cognitive demand of having raters verbalize their thoughts while scoring is not representative of rating procedures (Lumley, 2005), the other half of the raters provided scores without verbalizing their thoughts during their first listening. This was a timed condition, with a 7-second interval between speech samples (duration: 18 min).

Raters performed a second listening immediately after finishing their first set of ratings, consulting their scores. When the recording was paused, raters articulated what they remembered thinking about the rating process or their impressions of the speech. For half of the raters not in the verbal protocol condition described above, these delayed recalls were their only opportunity to comment on factors that had fed into their listening and scoring. However, the time lapse meant that the introspective reports were removed from their initial thought processes when rating (Ericsson & Simon, 1993). Finally, at the end of the session, all raters were interviewed about their scoring behavior, think-aloud experience, interpretations of the constructs, and perceived influences on their judgments. The interview data are not discussed in this article.

3.6 Rating scale normalization

Table 3 shows the equivalencies that we used to scale the 9-point scale down to a 5-point scale in preparation for data analysis. Isaacs & Thomson (2013) found that rater consistency was similar across scale length condition, the distributions of rating outcomes for each rated measure were virtually identical, and rater preference for using 9- versus 5-point scales was mixed, with no rater consensus achieved. Therefore, we pooled ratings across scale length condition using the normalized scales.

Table 3. Original and normalized scales for comprehensibility, accentedness, and fluency ratings

	Scale levels								
Original 5-point scales	1		2		3		4		5
Original 9-point scales	1	2	3	4	5	6	7	8	9
Normalized 9-point scales	1	1.5	2	2.5	3	3.5	4	4.5	5

3.7 Deriving discrete linguistic measures from the L2 speech samples

In order to examine the discrete pronunciation and fluency measures that most strongly relate to experienced teacher-raters' and novice raters' global ratings for the two learner groups, we obtained segmental, prosodic, and temporal measures from the speech. For speech segments, a phonetically-trained research assistant annotated orthographically transcribed recordings to indicate error locations and type, specifically vowel and consonant substitutions, deletions, and additions. When marking substitutions, the research assistant was told to ignore instances where a non-English sound was substituted for English in a way that did not impact intelligibility (e.g., a trilled 'r' in place of an English 'r' was acceptable, as were palatalized fricatives in place of English 'h'). The second author, a phonetician, then verified the annotated transcripts, noting any differences of opinion. He agreed with the assistant's assessment in 93% of cases. There was greatest agreement on consonantal errors (97%), with less agreement on vowels (88%), which are notoriously ambiguous (McAndrews & Thomson, 2007). After considering each discrepancy, when consensus was not possible, the second author's judgment stood. This only affected a few decisions related to vowels and one related to a consonant error. In most cases where there was disagreement, vowel productions were determined to be ambiguous and were subsequently accepted as correct. Previous studies have used blind randomized assessment of discrete speech tokens produced from a word list using a forced-choice decision task (e.g., Thomson & Isaacs, 2009). We did not feel that this approach was suitable for the current study, since unpredictable speech tokens arising in extemporaneous narratives were the focus rather than discrete items targeting specific sounds. In the final analysis, there were an average of 4.2 vowel errors (range: 0–10) and 4.6 consonant errors (range: 0–10), per 20 second L2 speech sample.

We computed ratios of correctly pronounced segments over segmental incidence, tabulated separately for vowels and consonants in content versus function words. We distinguished between these word types because Zielinski's (2008) in-depth analysis revealed little role for function words in intelligibility breakdowns. However, her study analyzed only three L2 learners' speech samples. Further, Munro and Derwing (2006) provided evidence supporting the functional load hypothesis in relation to comprehensibility, albeit with the potential confound that in their stimuli, high functional load errors solely occurred in content words, which, by definition, are more consequential for meaning than function words. Therefore, we examined error prevalence for vowels and consonants in content versus function words and related this to the mean L2 speech ratings. We also computed the percent of correctly pronounced segments in pruned content versus function words (i.e., with all disfluencies removed), with vowels and consonants counted separately.

Prosody was captured by eliciting three pronunciation experts' prosodic goodness ratings using 9-point scales (1 = extremely non native prosody; 9 = native like prosody) following Derwing, Rossiter, Munro, and Thomson (2004). The experts were L2 pronunciation researchers and teachers with phonetic training and at least 15 years' residence in the Canadian province where the speech samples had been collected. Cronbach's alpha was used to confirm high internal consistency (.90) for the resulting prosodic goodness ratings. Drawing on Derwing et al. (2006), we examined two temporal measures using Sound Studio 3: (1) speaking rate, operationalized as the total number of uttered syllables over speech sample duration, and (2) pruned syllables per second, operationalized as the proportion of uttered syllables per second with all disfluencies removed (e.g., self-repetitions, self-corrections). We used 400 milliseconds as the minimum threshold for counting silent pauses or fillers (see Derwing et al., 2004; Riggenbach, 1991).

3.8 Analysis of introspective reports

The verbal protocol and delayed recall data were orthographically transcribed and verified by the second researcher. Words with irregular pronunciation that raters had recalled or imitated from the speech samples were written with phonemic symbols or underlined for stress. To examine the linguistic aspects that experienced teacher-raters and novice raters reportedly attended to in Mandarin and Slavic language speakers' utterances, the first author inductively generated a coding scheme in an iterative process based on raters' verbatim comments. Twenty verbal protocols and 20 delayed recalls were subjected to coding and enumeration so that only one set of comments per rater was included for each speech sample (i.e., their first think-aloud opportunity). Coded categories and subcategories included: (1) segmental errors, identifying, where possible, error type (epenthesis, substitution, deletion) and whether vowels or consonants were implicated; (2) word pronunciation difficulty, in which raters expressed difficulty with or a pronunciation irregularity of a lexical item, but the error source could not be identified from the rater's comment; (3) word stress; (4) pitch, intonation, or voice quality (including pleasant/strange voice); (5) rhythm and linking (e.g., smooth/choppy speech); (6) pausing and other hesitation markers, specifying whether the comment pertained to filled or unfilled pauses where possible; and (7) speech rate or pacing (fast/reasonable vs. slow/halting delivery). Positive and negative comments about categories 3 through 7 were tallied separately. The coding scheme also captured general comments about the global rated measures (comprehensibility, accentedness, fluency), speakers' presumed personality attributes extrapolated from the speech (e.g., confidence), and rater processes or strategies. Comments about storytelling ability, grammatical use, syntactic complexity, and lexical appropriateness were not included in the coding scheme, although Derwing and Munro (1997) and Isaacs and Trofimovich (2012) have shown that comprehensibility pertains to more than simply pronunciation and fluency phenomena. Because raters were not constrained in the length of their introspective reports and we used a balanced design, the recording time or number of words uttered was not controlled for in subsequent analyses.

After obtaining the research team's feedback on the coding scheme, the following refinements were made. Pronoun errors were interpreted as grammatical rather than lexical errors, self-repetition was classified under the pausing/hesitations category, and stuttering fell under rhythm/linking. A second coder then applied the coding scheme to the data, recording frequencies separately for rater experience and speaker L1. Exact intercoder agreement was obtained 93% of the time for the main categories, with differences of opinion resolved through discussion. Discrepant codes were assigned, for example, when one coder interpreted "stops" to mean plosives, whereas a closer reading revealed that the rater was, in fact, referring to stops and starts. Comments about lexical retrieval difficulties resulting in disfluency or inadequate information produced, which were a source of coding inconsistency, were ultimately assigned the pausing/hesitation code, except for instances when the rater directly referred to slow speech or processing as being an issue, in which case speech rate/pacing was selected. In ambiguous cases when an error type could not be classified based on the rater's account, the audio recordings of the introspective reports were consulted to check the fidelity of the transcription and coding interpretation.

After finalizing the frequency counts, the five main coded categories that, together with subcategories, were most frequent in the data were submitted to log-linear analysis using SAS 9.4 GENMOD and CATMOD procedures. This yielded a crosstabulation of categorical variables using chi-square tests for statistical significance and maximum likelihood estimation (Stevens, 2009). All other statistical analyses were computed using SPSS 24.

4. Results

4.1 Preliminary analyses

Before addressing the research questions, we conducted three preliminary analyses. First, intraclass correlations for ratings of comprehensibility (.964), accentedness (.965), and fluency (.972) revealed high internal consistency. Next, an independent samples *t*-test, conducted to examine whether there were scoring differences for raters assigned to the verbal protocol versus delayed recall conditions, which was an artifact of the research design, revealed no significant differences, $t(38) = |.01–1.38|$, $p > .05$. Therefore, we pooled ratings across introspective report conditions and ran Pearson correlations between the three global rated measures. The moderate to strong associations in Table 4 suggest that these constructs are related yet distinct.

Table 4. Correlations between L2 comprehensibility, accentedness, and fluency ratings

Rated measures	1	2	3
1 Comprehensibility			
2 Accentedness	.71**		
3 Fluency	.65**	.61**	

* $p \leq .05$ ** $p \leq .01$, two-tailed

4.2 Rater experience and speaker L1 in relation to global ratings and discrete measures

A series of partially repeated measures ANOVAs were conducted, with speaker L1 a within-subjects' factor and rater experience a between-subjects factor. For comprehensibility ratings, we found a significant main effect for speakers' L1, $F(1,38) = 248.026$, $p < .001$, partial $\eta^2 = .867$, but not for rater experience. For accentedness ratings, we found significant main effects for speakers' L1, $F(1,38) = 233.156$, $p < .001$, partial $\eta^2 = .860$, but not for rater experience. For fluency ratings, we found a significant main effect for speakers' L1, $F(1{,}38) = 230.681$, $p = .<001$, partial $\eta^2 = .859$, but not for rater experience. There were no significant interaction effects.

In sum, there were no significant group differences in how experienced teacher-raters and novice raters scored all speakers, but pooled across raters, the speakers' L1 did affect comprehensibility, accentedness and fluency ratings. Slavic language speakers were rated as significantly more comprehensible, significantly less accented and significantly more fluent compared to Mandarin speakers. These findings are not surprising given the extremely strong Pearson correlations between mean ratings provided by the experienced teacher-raters and novice raters (see Figure 1). Figure 2 shows mean ratings by L1 background for the experienced teacher and novice groups combined.

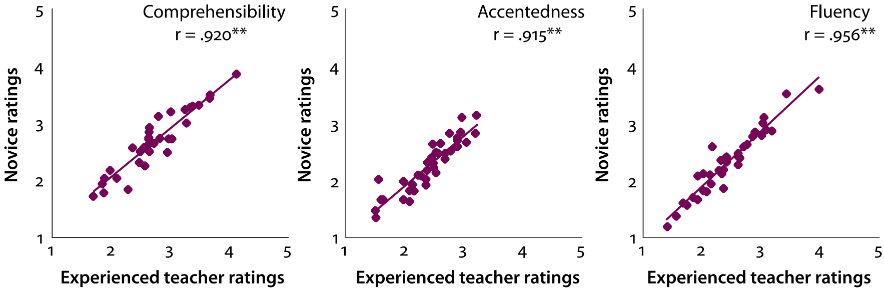


Figure 1. Scatterplots of mean experienced teacher raters' and novice raters' scores for each L2 speaker using normalized comprehensibility, accentedness and fluency scales

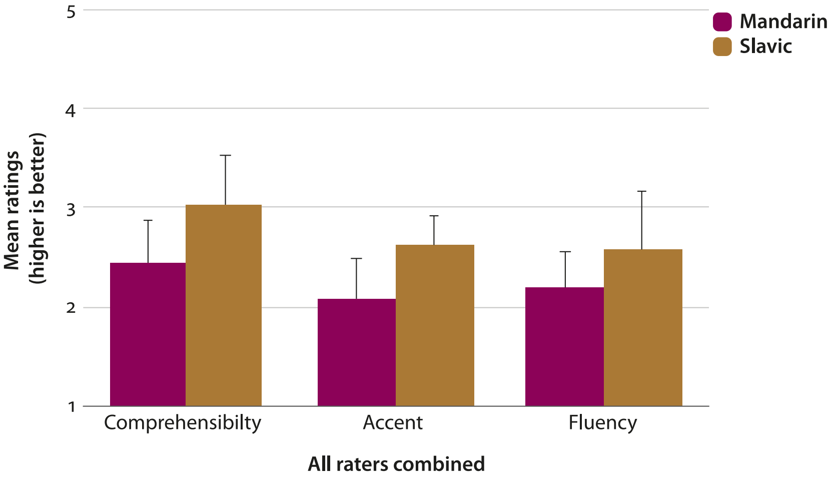


Figure 2. Mean comprehensibility (Comp.), accentedness (Acc) and fluency (Flu) ratings on the normalized scales by speakers' L1 background. Bars enclose ±1 *SD*

Next, we computed correlations between the three global rated measures pooled across all raters and segmental accuracy (in content and function words), prosodic goodness, and the temporal measures (pruned syllables/s and speaking rate). Results revealed a nearly perfect correlation between prosodic goodness and L2 comprehensibility ratings, *r*=.98. Strong correlations were also revealed between prosodic goodness and ratings of both fluency, *r*=.91, and accentedness, *r*=.83. The proportion of correctly pronounced segments in content words was moderately associated with ratings for accentedness, *r*=.58, comprehensibility, *r*=.55, and fluency, *r*=.36. However, in function words, there was a very weak to no relationship with any of the three global rated constructs. The ratio of segmental errors over segmental incidence for vowels, consonants, and both are presented in Figures 3 and 4 for comprehensibility and accentedness ratings, respectively. The correlation is slightly higher for vowel than consonant accuracy measures, particularly for accentedness. Finally, both temporal measures strongly correlated with fluency, with a moderate relationship with comprehensibility and a moderate to weak association with accentedness.

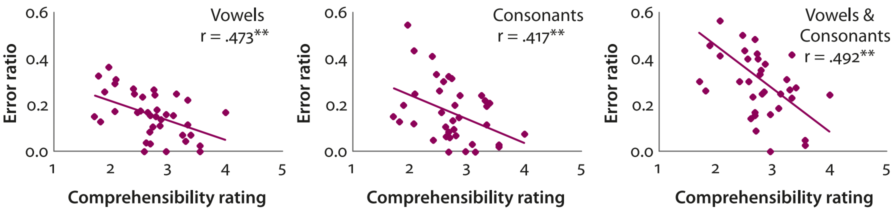


Figure 3. Ratio of segmental errors (vowels, consonants, or combined) to total errors in relation to comprehensibility ratings

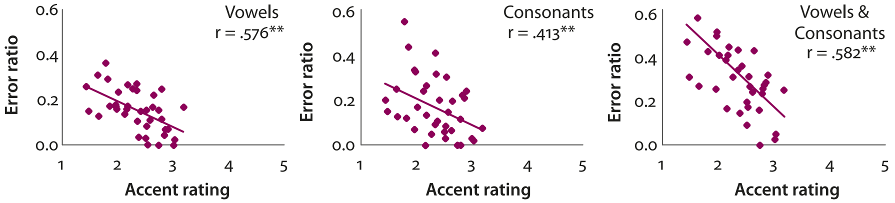


Figure 4. Ratio of segmental errors (vowels, consonants, or combined) to total errors in relation to accentedness ratings

We then broke these findings down by the two independent variables of interest. For experienced-teacher raters versus novice raters, the overall patterns of association were similar (see Table 5). However, the temporal measures were more strongly associated with novice than experienced raters' overall perceptual judgments, whereas prosodic goodness was more strongly related to experienced teachers' than novice raters' fluency judgments. Table 6 shows a much stronger relationship between the two temporal measures and both comprehensibility and accentedness ratings for the L1 Slavic compared to Mandarin speakers. This implies that the overall ratings of the Mandarins' speech productions are not captured as well by these measures.

Table 5. Correlations between mean L2 comprehensibility, accent, and fluency ratings, and discrete speech measures grouped by rater experience

	Comprehensibility		**Accentedness**		**Fluency**	
	Experienced	**Novice**	**Experienced**	**Novice**	**Experienced**	**Novice**
Pruned content word segmental accuracy	.55**	.52**	.56**	.57**	.38*	.33*
Pruned function word segmental accuracy	.21	.16	.28	.27	−.01	−.02
Prosodic goodness	.96**	.96**	.82**	.80*	.93**	.87**
Speaking rate	.54**	.55**	.32	.36*	.79**	.81*
Pruned syllables/s	.58**	.62**	.37*	.46*	.78**	.83**

* $p \leq .05$ ** $p \leq .01$, two-tailed

Table 6. Correlations between mean comprehensibility, accent, and fluency ratings, and discrete speech measures grouped by L1 background

	Comprehensibility		**Accentedness**		**Fluency**	
	Mandarin	**Slavic**	**Mandarin**	**Slavic**	**Mandarin**	**Slavic**
Content word segmental accuracy	.425	.379	.362	.551**	.214	.277
Function word segmental accuracy	.189	−.112	.145	.115	.041	−.240
Prosodic goodness	.976**	.975**	.760**	.807**	.839**	.953**
Speaking rate	.450	.756**	.254	.518**	.742**	.876**
Pruned syllables/s	.419	.797**	.174	.637**	.713**	.869**

* $p \leq .05$ ** $p \leq .01$, two-tailed

4.3 Analysis of the factors that raters reportedly take notice of when rating L2 speech

Having clarified the relationship between global L2 speech ratings and discrete measures in relation to rater experience and speaker L1, we sought to examine the factors to which experienced teacher-raters versus novice raters reportedly attend when rating Mandarin and Slavic language speakers' utterances (research question 2). Table 7 shows frequency counts of the coded comments and loglinear analysis results for the five main categories that were most frequent. Figures 5 and 6 show counts of coded categories or subcategories by experience and L1, respectively.

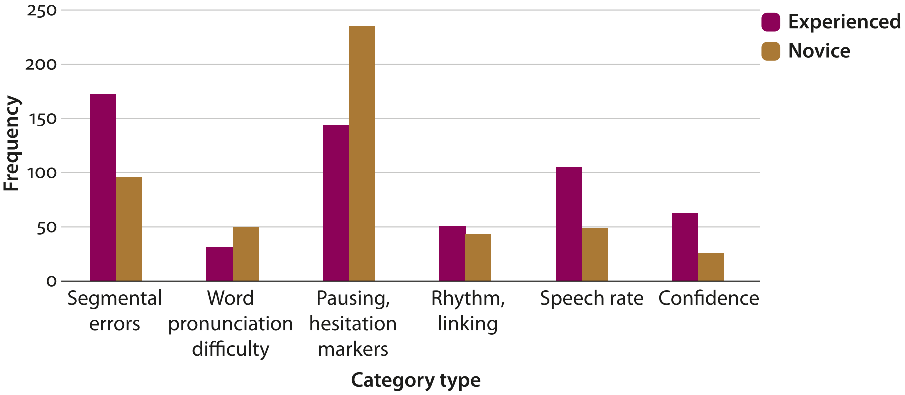


Figure 5. Frequency of coded comments by category type grouped by rater experience

Experienced teacher-raters' introspective reports were longer than those of novice raters, producing significantly more comments for all coded categories and subcategories. The exceptions to this were comments about pausing and "word pronunciation difficulty," in which a pronunciation irregularity was signaled in the comments but the specific error type could not be identified in the coding based on the rater's account (e.g., "mispronounced a couple of words that made the words incomprehensible"). This may be because novice raters observed little else about the speech or lacked the vocabulary with which to pinpoint other features. Conversely, experienced teacher-raters were more precisely able to articulate the error source or more frequently imitated a lexical item such that the error type could be identified. Experienced teacher-raters may also have been more invested in the task than novice raters, which could partially account for their lengthier verbalizations. Overall, comment frequencies about rhythm and linking revealed no rater group differences. However, experienced teacher-raters made significantly more positive comments about these elements than novice

Table 7. Frequencies of coded comments and loglinear analysis[a] by rater experience and speaker L1

	Mandarin experienced	Slavic experienced	Mandarin novice	Slavic novice
Total segmental errors comments	**109**	**63**	**53**	**43**
Experience: χ^2 (1,39) = 20.95, p < .0001				
L1: χ^2 = (1,39) = 11.53, p = .0007				
Total vowel errors	48	24	22	18
Experience: χ^2 = 8.88, p = .003				
L1: χ^2 = 6.85, p = .009				
Epenthesis	15	4	4	3
Substitution	26	15	12	8
Deletion	2	–	1	–
Error source unclear	5	5	5	7
Total consonant errors	55	34	26	18
Experience: χ^2 = 14.61, p = .0001				
L1: χ^2 = 6.22, p = .0126				
Epenthesis	8	5	5	2
Substitution	27	20	14	16
Deletion	12	1	4	–
Error source unclear	8	8	3	–
Segmental error unclassifiable	6	5	5	7
Word pronunciation difficulty (unclassifiable pronunciation errors)	**26**	**5**	**42**	**8**
Experience: χ^2 = –4.78, p = .037				
L1: χ^2 = 29.88, p < .0001				
Total rhythm/linking comments	**26**	**25**	**20**	**18**
Good rhythm/linking	12	18	4	11
Experience: χ^2 = 4.80, p = .0284				
Poor rhythm/ linking	14	7	16	7
L1: χ^2 = 5.54, p = .0185				
Total pausing-related comments	**60**	**84**	**117**	**118**
Experience: χ^2 = –4.89, p < .0001				
Silent pauses	7	12	22	19
Filled pauses	12	20	30	35
Disfluency source unclear	41	52	65	64

Table 7. *(continued)*

	Mandarin experienced	Slavic experienced	Mandarin novice	Slavic novice
Total speech rate comments	**50**	**55**	**21**	**28**
Experience: $\chi^2 = 19.41$, $p < .0001$				
Fast/reasonable pace	9	15	2	8
Experience: $\chi^2 = 5.4$, $p = .020$				
L1: $\chi^2 = 4.05$, $p = .044$				
Slow pace	41	40	19	20
Experience: $\chi^2 = 14.06$, $p < .001$				
Total comments about confidence	**23**	**40**	**12**	**14**
Experience: $\chi^2 = 14.42$, $p < .0001$				
L1: $\chi^2 = 3.99$, $p = .0457$				
Speaker confident	16	29	7	9
Experience: $\chi^2 = 12.62$, $p = .0004$				
Speaker unconfident	7	11	5	5

a Only statistically significant main effects are shown for the chi-square results ($p \leq .05$). No significant interaction effects were detected.

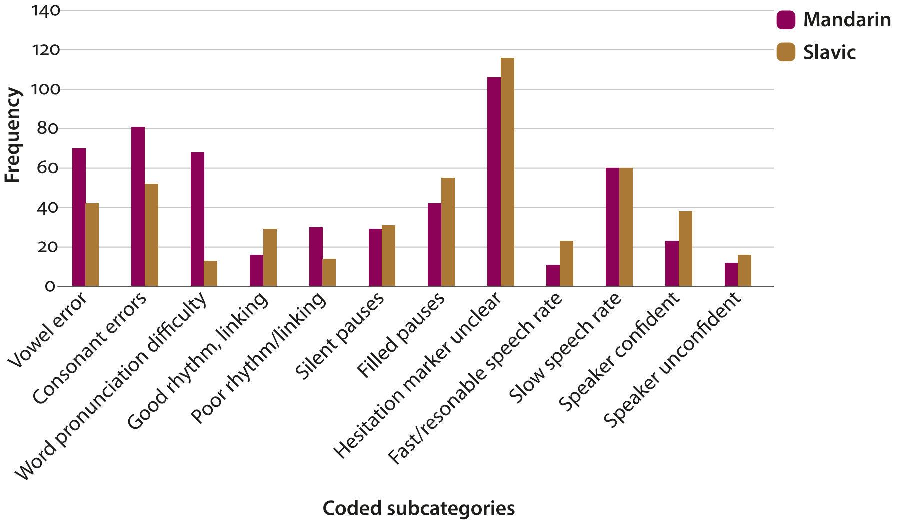


Figure 6. Frequency of coded comments for subcategories grouped by speakers' L1

raters. They also commented more about how confident the speaker sounded. Frequency counts for word stress and pitch/intonation/voice were too low to be included in the loglinear analysis.

Mandarin speakers received more comments about segmental errors than Slavic language speakers, with higher frequency counts for consonants than vowels in the contingency table. There was a main effect for L1 for both vowels and consonants, with a larger effect size for vowels. This could suggest that the vowel errors that raters pinpointed for Mandarin speakers may have been more salient or consequential compared to the more numerous consonant errors identified. Raters also appeared to struggle with word pronunciation when listening to Mandarin compared to Slavic language speakers and provided more negative comments on rhythm/linking for Mandarins. However, pausing was commented on significantly more frequently for Slavic language speakers. Raters also noted a fast/reasonable speech rate more often for Slavic language speakers, although comments about slow paced speech and pausing were nonsignificant across groups. Finally, more comments extrapolating speakers' confidence levels from the speech samples were made for L1 Slavic than Mandarin speakers.

5. Discussion

5.1 Rater experience

This mixed methods study examined one rater characteristic (experience) and one speaker variable (L1) in relation to L2 comprehensibility, accentedness, fluency ratings, how segmental, temporal, and prosodic measures relate to these constructs, and raters' reported influences when scoring the speech. Our first main finding that experienced teacher-raters' and novice raters' scores were not significantly different echoes Bongaerts, van Summeren, Planken, and Schils' (1997) nonsignificant result for accentedness. However, it contradicts both Thompson (1991), who found that experienced teacher-raters were harsher judges than novice raters for accentedness, and Rossiter (2009), who found that experienced teacher-raters were more lenient than novice raters for fluency. None of these studies examined comprehensibility, accentedness, and fluency together. A methodological explanation for these inconsistent findings across studies includes differences in how experienced and novice raters were operationalized, L2 speaker characteristics (e.g., L1 background, L2 proficiency), the speaking task(s) used, rater characteristics (e.g., accent familiarity), the rating scales used, the way that rater severity was computed, and statistical power. A systematic review or meta-analysis synthesizing the rater experience variable could help clarify the strength of the evidence and provide further methodological considerations.

Experienced teacher-raters' and novice raters' mean comprehensibility, accentendness, and fluency ratings were strongly correlated with the pronunciation experts' pooled goodness-of-prosody ratings, with a near perfect correlation for comprehensibility. This finding is consistent with research emphasizing the importance of prosodic features for comprehensibility (Isaacs & Trofimovich, 2012; Saito, Trofimovich, & Isaacs, 2016) and, for some L2 learners, intelligibility (Derwing & Munro, 1997; Hahn, 2004). However, two limitations need to be acknowledged. First, we did not apply a low pass filter for prosodic goodness ratings, which would have isolated prosodic phenomena and removed the distraction of segmental and morphosyntactic errors for the expert raters (Derwing & Munro, 1997). Therefore, the strength of association between prosodic goodness, comprehensibility, and other measures in this study should be treated with caution. Another limitation is that the more objective measure of intelligibility, which, by definition, captures actual rather than perceived listener understanding, was not examined here.

Next, we found that researcher-coded segmental accuracy ratios were moderately related to raters' mean L2 accentedness and comprehensibility ratings, with a larger role for vowels than consonants, particularly for accentedness. This result, especially for comprehensibility, which applied linguists widely consider an appropriate goal for L2 pronunciation teaching and assessment (Isaacs & Harding, 2017), implies that segments should not be ipso facto discounted in favor only of prosodic instruction. This view is consistent with previous research demonstrating a role for high functional load segmental errors in impeding comprehensibility (Munro & Derwing, 2006), distinguishing between different L2 speaking levels (Kang & Moran, 2014), and detracting from some L1 groups' comprehensibility (Suzukida & Saito, 2019).

Whereas accurately pronounced pruned segments in content words were moderately correlated for both experienced teacher-raters' and novice raters' L2 accentedness and comprehensibility ratings, in function words, this measure had a nonsignificant relationship with the global rated measures. This suggests that Zielinski's (2008) finding that function words are rarely implicated in intelligibility breakdowns extends to comprehensibility. Put simply, segmental errors in content words are a more robust measure (and more consistent with the meaning-laden nature of comprehensibility) than segmental error measures that also include function words. Consequently, we suggest that function words be removed from segmental accuracy measures or, alternatively, that functional load or some other way of gauging error locus or gravity be taken into account.

Correlations between the global rated measures and two temporal measures (pruned syllables per/s and speaking rate) were marginally higher for novice than experienced teacher-raters. This finding roughly aligns with results from the introspective reports. Although experienced teacher-raters verbalized their thoughts more fully than novice raters, the sole category where the frequency of novice raters' comments exceeded that of experienced teacher-raters was for pausing. This may be because pausing was particularly salient and disruptive for novice raters. Alternatively, pausing may have been easier for them to discuss than other linguistic phenomena, for which they lacked the vocabulary, or may have served as the default option when they had little else to say. As for experienced teacher-raters, previous research has shown that that even teachers who have served as accredited examiners or textbook authors can have difficulty with pronunciation-related terminology (Foote, Isaacs, & Trofimovich, 2013; Isaacs, Trofimovich, Yu, & Chereau, 2015). This finding did not apply uniformly to the experienced teacher-raters in our study, with nearly a third reporting pronunciation training. Whereas some used technical terms in their introspective reports to refer to pronunciation and fluency phenomena (e.g., "sibilants," "semivowel," "primary stress"), others used more colloquial language (e.g., "mangles vowel sounds," "r's… swallowed," "putting noise in between what he's saying" for filled pauses). Such variability within the experienced teacher group is noteworthy. However, there were still overall differences with the novice group in terms of talk quantity, linguistic features emphasized, and likely pronunciation literacy levels.

The only other coded category where the frequency of comments for novice raters was higher than for experienced teacher-raters was for word pronunciation difficulty, designating an unclassifiable error type. This suggests that novice raters may have struggled to recall or articulate the source of a pronunciation difficulty that they had noticed. Such explanations are speculative, and it would be useful to examine raters' accounts of their observations and processes using the follow-up interviews. Similarly, as most existing L2 pronunciation and fluency research on rater experience has been primarily quantitative (e.g., Rossiter, 2009; Saito, Trofimovich, Isaacs, & Webb, 2017), future studies could triangulate statistical findings with qualitative data to better understand rater orientations.

Although we have emphasized differences between experienced teacher-raters and novice raters above, the correlations patterns between discrete linguistic features and global speech measures was similar, with correlations coefficients at most only .06 different between groups. These values were less divergent than in Rossiter's (2009) L2 fluency development study, suggesting the need for further investigation. Future research could also compare ESL teachers' scoring behaviour and perspectives with those of people who do not spend their working days with L2 speakers but, nonetheless, interact with them regularly (e.g., as work colleagues).

5.2 Speaker L1 background

The Slavic language speakers were rated significantly higher than their Mandarin peers for comprehensibility, accentedness, and fluency ratings, despite both rater groups reporting significantly more exposure to Mandarin- than Russian-accented English. This familiarity effect would likely have advantaged the Mandarin speakers (Browne & Fulcher, 2017), but they were still judged more harshly. Bongaerts, Mennen, & van der Slik (2000) suggest that such results may be partially explained by the phonological distance between learners' L1 and L2. Despite being potentially more familiar to listeners, Mandarin accented English may contain more divergences from English than Slavic accented English. For example, with a few exceptions, Mandarin disallows coda consonants. Transferred to English, dropping coda consonants and/or vowel insertion could have a strong effect on Mandarin learners' comprehensibility relative to Slavic language speakers' utterances, which would not contain the same error types (McAndrews & Thomson, 2017). Ultimately, familiarity with a particular accent cannot, on its own, predict how accented or comprehensible speech in that accent is to listeners. Phonological distance is also known to play a role (Bradlow, Clopper, Smiljanic, & Walter, 2010). While Bradlow et al. (2010) did not explicitly measure the phonological distance between Russian/Ukrainian and English and Mandarin an English, they did examine phonological distances between other Slavic languages (Slovene and Croatian) and English and between Cantonese and English. Their evaluation concluded that the Slavic languages are phonologically much more similar to English than Cantonese is to English.

The relationship between the temporal measures and listeners' L2 comprehensibility and accentedness and fluency ratings was moderate for Slavic language speakers, whereas for Mandarin speakers there was a significant correlation between temporal measures and fluency ratings, but not with comprehensibility and accentedness ratings. For prosodic goodness, all correlations were strong, but the association was stronger for the Slavic language than Mandarin speaking group. Finally, for content word segmental accuracy, the sole significant relationship was for Slavic language speakers' accentedness ratings. This suggests that raters may have been preoccupied by extraneous features of Mandarins' speech not accounted for by the segmental, prosodic, and temporal measures examined. For example, none of the measures captured morphosyntax or task execution, which could have been subject to L1 differences. It could also be that raters were overwhelmed by the amount of divergence of Mandarin learners' speech due to its typological dissimilarity with English, such that the linguistic measures were less related to the global rated constructs than for Slavic language speakers. Further research could incorporate a wider range of linguistic measures and gauge their sensitivity in capturing the variance in L2 speaking performances for different L1 groups. Saito, Webb, Trofimovich, & Isaacs (2016), for example, focused on a set of lexical measures in relation to L2 comprehensibility and accentedness ratings. More research investigating macro-level discourse measures using longer speech samples would also be useful.

Although not statistically significant, the association between comprehensibility and content word segmental accuracy was higher for Mandarin than for Slavic language speakers. The loglinear analysis revealed significant main effects for word pronunciation difficulty and segmental errors, with frequencies of coded comments higher for Mandarin than Slavic language speakers. Although consonant-related comments were more numerous for both L1 groups, the effect size was higher for vowels, in line with the correlation analysis in Figures 3 and 4. This finding supports previous pronunciation research on L1 effects emphasizing the contribution of segmental errors to Mandarin speakers' comprehensibility (Crowther, Trofimovich, Saito, & Isaacs, 2015).

There were no significant L1 group differences for the frequency of rater comments about disfluency markers by L1. However, pure frequency counts of coded comments suggest that filled pauses may have been more perceptually salient for Slavic than Mandarins language speakers. It may be that L1 influence in the articulation of fillers was more noticeable for Slavic language speakers (de Boer & Heeren, 2019), although formant frequencies were not obtained and filled pause duration only indirectly factored into the pruned syllables measure. Whereas significantly more comments were generated about Mandarin speakers' poor rhythm or linking in the introspective reports, Slavic language speakers received significantly more comments about having fast or reasonably paced speech. Raters also commented more about Slavic language speakers' confidence, although the number of positively or negatively coded comments did not translate into significant L1 differences.

5.3 Concluding remarks

This study moves beyond most existing L2 pronunciation and fluency research by examining not only linguistic measures drawn from L2 speech samples, but also raters' accounts of the linguistic features they reportedly pay attention to when scoring L2 speech. Ensuring that raters interpret the focal constructs in the same way while taking into account construct-relevant features is important for construct validity, with implications for rater screening and training in research and assessment settings. We acknowledge that examining the frequency of raters' comments, which they are conscious of and willing/able to articulate, is an imperfect proxy of what they are actually attending to (Ericsson & Simon, 1993).

Further, listeners may not understand their own analytic processes (Munro, 2018), and post-hoc reporting is prone to rationalization and face-saving strategies. Because methods for examining what goes on in raters' minds in light of their interaction with L2 speaking performances, tasks, and scoring systems are indirect, research evidence needs to be triangulated using multiple data sources to paint a more complete picture. In addition, moving beyond observational studies to examine causal relationships between linguistic deviations and ratings using experimental or quasi-experimental designs would be desirable.

We suggest that L2 pronunciation research would benefit from greater exploration of rater processes. Most existing studies focus on which linguistic measures/dimensions account for the variance in global L2 speech ratings without examining how raters arrive at their scoring decisions (e.g., Saito et al., 2016). Future research could incorporate an eye-tracking component to examine rater fixations on different scale bands, be they numerical scales or more elaborated descriptors. The resulting evidence could then be triangulated with other data sources (e.g., stimulated recalls, interviews, ratings). In sum, we highlight here the importance of investigating individual and group differences in listeners' approaches to rating. Such research could elucidate key methodological issues in running experiments or operational L2 assessments with a pronunciation or fluency component (e.g., O'Brien, 2016 found no scale sequencing effects).

Finally, this study has focused on linguistic measures derived from L2 speech and raters' introspective reports. However, variables extraneous to the properties of L2 speaking performances may also be reflected in ratings, posing problems for score interpretation. For example, so-called rater effects, such as listeners' exposure to or attitudes toward L2 accented speech, could influence their scoring decisions (e.g., Winke, Gass, & Myford, 2013). However, negative rater judgments should not automatically be dismissed as prejudicial (Munro, 2018). Future research should ideally examine rater characteristics or orientations that could threaten the validity of the L2 abilities being measured within the same research program as construct-relevant factors.

Acknowledgments

Thanks are due to Garrett Byrne, Daniel Clement, Marc Garellek, Joseph Hartfeil, Donna Pearce, and Monika Spak, who assisted with data transcription or coding, Carolyn Turner and Pavel Trofimovich for their input on the content of the paper, Tracey Derwing and Murray Munro for sharing their speaking prompt, the teachers and administrators who facilitated participant recruitment, and our participants, who spoke candidly about their impressions. We are also grateful to the editors for their insightful comments on previous versions of this article.

References

Bannigan, K., & Watson, R. (2009). Reliability and validity in a nutshell. *Journal of Clinical Nursing*, 18(23), 3237–3243. https://doi.org/10.1111/j.1365-2702.2009.02939.x

Bongaerts, T., Mennen, S., & van der Slik, F. (2000). Authenticity of pronunciation in naturalistic second language acquisition: The case of very advanced late learners of Dutch as a second language. *Studia Linguistica*, 54(2), 298–308. https://doi.org/10.1111/1467-9582.00069

Bongaerts, T., van Summeren, C., Planken, B., & Schils, E. (1997). Age and ultimate attainment in the pronunciation of a foreign language. *Studies in Second Language Acquisition*, 19(4), 447–465. https://doi.org/10.1017/S0272263197004026

Bradlow, A., Clopper, C., Smiljanic, R., & Walter, M.A. (2010). A perceptual phonetic similarity space for languages: Evidence from five native language listener groups. *Speech Communication*, 52(11–12), 930–942. https://doi.org/10.1016/j.specom.2010.06.003

Browne, K., & Fulcher, G. (2017). Pronunciation and intelligibility in assessing spoken fluency. In T. Isaacs & P. Trofimovich (Eds.), *Second language pronunciation: Interdisciplinary perspectives* (pp. 37–53). Bristol, UK: Multilingual Matters.

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing*, 12(1), 62–70. https://doi.org/10.1177/026553229501200102

Creswell, J.W., & Plano Clark, V.L. (2017). *Designing and conducting mixed methods research* (3rd ed.). Thousand Oaks, CA: SAGE.

Crowther, D., Trofimovich, P., Saito, K., & Isaacs, T. (2015). Second language comprehensibility revisited: Investigating the effects of learner background. *TESOL Quarterly*, 49(4), 814–837. https://doi.org/10.1002/tesq.203

de Boer, M., & Heeren, W. (2019). The speaker-specificity of filled pauses: A cross-linguistic study. *Proceedings of the International Congress of Phonetic Sciences (ICPhS) 2019* (pp. 607–611). Melbourne, Australia: Australasian Speech Science and Technology Association.

Derwing, T.M., & Munro, M.J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, 19(1), 1–16. https://doi.org/10.1017/S0272263197001010

Derwing, T.M., & Munro, M.J. (2009). Comprehensibility as a factor in listener interaction preferences: Implications for the workplace. *Canadian Modern Language Review*, 66(2), 181–202. https://doi.org/10.3138/cmlr.66.2.181

Derwing, T.M., & Munro, M.J. (2013). The development of L2 oral language skills in two L1 groups: A 7-year study. *Language Learning*, 63(2), 163–185. https://doi.org/10.1111/lang.12000

Derwing, T.M., Rossiter, M.J., Munro, M.J., & Thomson, R.I. (2004). Second language fluency: Judgments on different tasks. *Language Learning*, 54(4), 665–679. https://doi.org/10.1111/j.1467-9922.2004.00282.x

Derwing, T.M., Thomson, R.I., & Munro, M.J. (2006). English pronunciation and fluency development in Mandarin and Slavic speakers. *System*, 34(2), 183–193. https://doi.org/10.1016/j.system.2006.01.005

Douglas, D. (1994). Quantity and quality in speaking test performance. *Language Testing*, 11(1), 125–144. https://doi.org/10.1177/026553229401100203

Ericsson, K.A., & Simon, H.A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/5657.001.0001

Foote, J.A., Isaacs, T., & Trofimovich, P. (2013, June 3–5). *Developing a teacher-friendly assessment tool for L2 comprehensibility*. Canadian Association of Applied Linguistics (ACLA/CAAL) conference, Calgary, AB.

Foote, J.A., & Trofimovich, P. (2018). Is it because of my language background? A study of language background influence on comprehensibility judgments. *Canadian Modern Language Review*, 74(2), 253–278. https://doi.org/10.3138/cmlr.2017-0011

Gass, S.M., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah, NJ: Lawrence Erlbaum.

Hahn, L.D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly*, 38(2), 201–233. https://doi.org/10.2307/3588378

Isaacs, T., & Harding, L. (2017). Research timeline: Pronunciation assessment. *Language Teaching*, 50(3), 347–366. https://doi.org/10.1017/S0261444817000118

Isaacs, T., & Thomson, R.I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, 10(2), 135–159. https://doi.org/10.1080/15434303.2013.769545

Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition*, 34(3), 475–505. https://doi.org/10.1017/S0272263112000150

Isaacs, T., Trofimovich, P., Yu, G., & Chereau, B.M. (2015). Examining the linguistic aspects of speech that most efficiently discriminate between upper levels of the revised IELTS pronunciation scale. *IELTS research reports online series*, 4.

Kang, O., & Moran, M. (2014). Functional loads of pronunciation features in nonnative speakers' oral assessment. *TESOL Quarterly*, 48(1), 176–187. https://doi.org/10.1002/tesq.152

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, 40(3), 387–417. https://doi.org/10.1111/j.1467-1770.1990.tb00669.x

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Frankfurt: Peter Lang.

McAndrews, M.M., & Thomson, R.I. (2017). Establishing an empirical basis for priorities in pronunciation teaching. *Journal of Second Language Pronunciation*, 3(2), 267–287. https://doi.org/10.1075/jslp.3.2.05mca

Munro, M.J. (2018). Dimensions of pronunciation. In O. Kang, R. Thomson, & J. Murphy. *The Routledge handbook of contemporary English pronunciation* (pp. 413–431). New York: Routledge.

Munro, M.J., & Derwing, T.M. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System*, 34(4), 520–531. https://doi.org/10.1016/j.system.2006.09.004

O'Brien, M.G. (2016). Methodological choices in rating speech samples. *Studies in Second Language Acquisition*, 38(3), 587–605. https://doi.org/10.1017/S0272263115000418

Pawlikowska-Smith, G. (2000). *Canadian Language Benchmarks 2000: Theoretical framework*. Ottawa, ON: Centre for Canadian Language Benchmarks.

Rajadurai, J. (2007). Intelligibility studies: A consideration of empirical and ideological issues. *World Englishes*, 26(1), 87–98. https://doi.org/10.1111/j.1467-971X.2007.00490.x

Riggenbach, H. (1991). Toward an understanding of fluency: A microanalysis of non-native speaker conversations. *Discourse Processes*, 14(4), 423–441. https://doi.org/10.1080/01638539109544795

Rose, H., & Galloway, N. (2019). *Global Englishes for language teaching*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316678343

Rossiter, M.J. (2009). Perceptions of L2 fluency by native and non-native speakers of English. *Canadian Modern Language Review*, 65(3), 395–412. https://doi.org/10.3138/cmlr.65.3.395

Saito, K., Trofimovich, P., & Isaacs, T. (2016). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics*, 37(2), 217–240. https://doi.org/10.1017/S0142716414000502

Saito, K., Trofimovich, P., Isaacs, T., & Webb, S. (2017). Re-examining phonological and lexical correlates of second language comprehensibility: The role of rater experience. In T. Isaacs & P. Trofimovich (Eds.), *Second language pronunciation assessment: Interdisciplinary perspectives* (pp. 131–146). Bristol, UK: Multilingual Matters.

Saito, K., Webb, S., Trofimovich, P., & Isaacs, T. (2016). Lexical correlates of comprehensibility versus accentedness in second language speech. *Bilingualism: Language and Cognition*, 19(3), 597–609. https://doi.org/10.1017/S1366728915000255

Schiavetti, N. (1992). Scaling procedures for the measurement of speech intelligibility. In R.D. Kent (Ed.), *Intelligibility in speech disorders* (pp. 11–34). Amsterdam: John Benjamins. https://doi.org/10.1075/sspcl.1.02sch

Stevens, J.P. (2009). *Applied multivariate statistics for the social sciences* (5th ed.). New York: Taylor & Francis.

Suzukida, Y., & Saito, K. (2019). Which segmental features matter for successful L2 comprehensibility? Revisiting and generalizing the pedagogical value of the Functional Load principle. *Language Teaching Research*. Advance online publication. https://doi.org/10.1177/1362168819858246

Thomson, R.I., & Isaacs, T. (2009). Within-category variation in L2 English vowel learning. *Canadian Acoustics*, 37, 138–139.

Thompson, I. (1991). Foreign accents revisited: The English pronunciation of Russian immigrants. *Language Learning*, 41(2), 177–204. https://doi.org/10.1111/j.1467-1770.1991.tb00683.x

Upshur, J.A., & Turner, C.E. (1999). Systematic effects in the rating of second-language speaking ability: Test method and learner discourse. *Language Testing*, 16(1), 82–111. https://doi.org/10.1177/026553229901600105

Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, 30(2), 231–252. https://doi.org/10.1177/0265532212456968

Zielinski, B.W. (2008). The listener: No longer the silent partner in reduced intelligibility. *System*, 36(1), 69–84. https://doi.org/10.1016/j.system.2007.11.004

Second language comprehensibility as a dynamic construct*

Pavel Trofimovich[1], Charles L. Nagle[2],
Mary Grantham O'Brien[3], Sara Kennedy[1], Kym Taylor Reid[1]
and Lauren Strachan[1]
[1] Concordia University | [2] Iowa State University | [3] University of Calgary

This study examined longitudinal changes in second language (L2) interlocutors' mutual comprehensibility ratings (perceived ease of understanding speech), targeting comprehensibility as a dynamic, time-varying, interaction-centered construct. In a repeated-measures, within-participants design, 20 pairs of L2 English university students from different language backgrounds engaged in three collaborative and interactive tasks over 17 minutes, rating their partner's comprehensibility at 2–3 minute intervals using 100-millimeter scales (seven ratings per interlocutor). Mutual comprehensibility ratings followed a U-shaped function over time, with comprehensibility (initially perceived to be high) being affected by task complexity but then reaching high levels by the end of the interaction. The interlocutors' ratings also became more similar to each other early on and remained aligned throughout the interaction. These findings demonstrate the dynamic nature of comprehensibility between L2 interlocutors and suggest the need for L2 comprehensibility research to account for the effects of interaction, task, and time on comprehensibility measurements.

Keywords: comprehensibility, pronunciation, dynamic, interaction, processing fluency, second language

1. Introduction

In their seminal work published 25 years ago, Munro and Derwing (1995) showed that intelligibility and comprehensibility of second language (L2) speech were

* An earlier version of this article was published as part of a special issue of the *Journal of Second Language Pronunciation* 6(3). https://doi.org/10.1075/jslp.20003.tro

https://doi.org/10.1075/bct.121.08tro

related yet partially independent constructs. In Derwing and Munro's framework, intelligibility is defined as "the extent to which a speaker's message is actually understood by a listener" (Munro & Derwing, 1995, p. 76), and comprehensibility refers to listeners' "judgments on a rating scale of how difficult or easy an utterance is to understand" (Derwing & Munro, 1997, p. 2). The constructs are partially independent because listeners who transcribe L2 utterances (which is a typical measure of intelligibility) nearly perfectly may nevertheless rate the same utterances as hard to understand. As a measure of ease or difficulty of understanding speech, comprehensibility has emerged as a key construct in empirical work focusing on linguistic, cognitive, and social variables associated with speech that is understandable to the listener (Derwing & Munro, 2015). However, in nearly all previous research, comprehensibility has been examined through one-time ratings (after speaking is completed) in monologic tasks (such as picture descriptions) and by listeners evaluating audio recordings only (without seeing speakers). To extend prior research on comprehensibility, we set out to provide a time-sensitive comprehensibility profile by focusing on comprehensibility within interaction, through ratings elicited from L2 speakers themselves as they perform communicative tasks.

2. Background literature

2.1 Comprehensibility: A measure of understanding

Typically assessed through listeners' transcriptions of speech content, intelligibility is often regarded as the gold standard for evaluating listener comprehension (Derwing & Munro, 2015). However, scalar ratings of comprehensibility are a useful measure of listener understanding in many contexts. To begin with, comprehensibility ratings are practical and intuitive, and they can be elicited and scored easily using speech samples featuring the same content. In contrast, intelligibility measures require tasks with unique speech content for each instance when intelligibility is measured (to avoid greater intelligibility for content that is repeated to listeners) and comparatively more time for listeners to complete the tasks. Comprehensibility ratings are also reliable across listeners, meaning that listeners generally agree with each other regardless of how comprehensibility is measured – through Likert-type scales (Munro & Derwing, 1995), sliding scales (Saito et al., 2017), or direct magnitude estimation (Munro, 2018). Most importantly, although intelligibility and comprehensibility are partially independent, comprehensibility ratings provide a reasonable estimate of listeners' actual understanding of speech (Sheppard et al., 2017). For instance, Munro and Derwing (1995) reported substantial overlap between these dimensions, with correlation coefficients approaching .90, although the magnitude of this link might vary for different speakers and listeners (Matsuura et al., 1999).

Besides being a practical measure of understanding, comprehensibility ratings are also shaped by the linguistic content of speech, which makes comprehensibility a useful metric to understand how various linguistic dimensions in the speaker's speech impact the listener. In their initial work, Munro and Derwing (1995) found associations between listeners' comprehensibility ratings and several linguistic measures derived from the speech being evaluated, including phonemic substitutions, intonation accuracy, and morphosyntactic errors. More recent work has revealed two constellations of linguistic dimensions relevant to comprehensibility: pronunciation (individual segments, prosody, fluency) and lexicogrammar (variety and richness of vocabulary, accuracy and complexity of grammar). The exact combinations of linguistic dimensions feeding into listeners' judgments of comprehensibility can depend on the linguistic background of the speaker and on the speaking task (Crowther et al., 2018), but the general finding has been consistent. Many measures at the level of segments, prosody, fluency, grammar, and discourse have been linked to listeners' ratings of L2 comprehensibility in multiple languages (Bergeron & Trofimovich, 2017; Crowther et al., 2015a; O'Brien, 2014).

2.2 Comprehensibility: An index of processing fluency

Comprehensibility ratings can also be conceptualized in a broader sense, as a measure capturing listeners' processing fluency, which refers to a person's subjective experience of the ease or difficulty with which information is processed (Reber & Greifeneder, 2017). A key aspect of processing fluency which cuts across various social and psychological domains is that people appraise and respond to various situations based on the perceived difficulty they report while processing a stimulus (e.g., text, image, sound), which may or may not reflect their actual experience with that stimulus. For instance, statements attributed to people whose names are harder to pronounce are considered less trustworthy (Newman et al., 2014), regardless of the actual content of the statements. Similarly, readers exposed to a text printed in a difficult to read font react more negatively to the reading than those who read the same text in an easy to read font, despite having similar text comprehension (Sanchez & Jaeger, 2015; Song & Schwarz, 2008). These findings are strikingly similar to Munro and Derwing's (1995) observation that comprehensibility might be rated differently for speech that is perfectly intelligible, implying that listeners' reactions to speech might be linked not to actual understanding (intelligibility) but to comprehensibility.

There is indeed growing evidence that comprehensibility (as a metric of processing fluency) captures important decisions for listeners. For instance, in social-psychological research on listeners' attitudes, speakers whom listeners perceived as hard to understand were downgraded in listeners' affective and attitudinal evaluations. Such speakers were ascribed negative emotions of annoyance and irritation and labelled less intelligent and successful (Dragojevic et al., 2017). Similarly, in a study focusing on online learning, when students evaluated an instructional video narrated by the instructor who was hard to understand, they downgraded the instructor in their evaluations, expressing negative attitudes towards online coursework and evaluating video content as more difficult, even though students' actual understanding of the video was not compromised (Sanchez & Khan, 2016). In fact, a comprehensibility scale akin to that used in L2 speech research has now been validated as part of a five-item processing fluency measure, and this measure appears to explain various human judgments (truthfulness, preference, perceived risk), all attributed to processing fluency in prior literature (Graf et al., 2018).

2.3 A dynamic approach to comprehensibility

Speaking and listening are dynamic acts whose properties fluctuate over time, yet comprehensibility has rarely been framed as a dynamic, variable process. Speakers generally alternate between periods of fluent and disfluent speech, with such temporal cycles recurring every 10–30 seconds (Pakhomov et al., 2011). And listeners must continuously adapt their comprehension to process varying levels of accuracy, complexity, and fluency to interpret the speaker's message within an emergent discourse structure (Kuperberg & Jaeger, 2016). Conversation is an inherently social process regarded as a joint, co-constructed activity between interlocutors (Brennan et al., 2018). Based on theoretical views that posit tight coordination between interlocutors (Garrod et al., 2018), comprehensibility could be characterized by variability both within and across interlocutors and could involve a continuous, dynamic adaptation of the interlocutors to each other, with comprehensibility sensitive to both global influences (e.g., time on task, task difficulty) and local issues (e.g., disfluency, error).

Nagle et al. (2019) recently explored whether comprehensibility can be construed as dynamic, examining how raters assign ratings in real time. In this study, 24 Spanish-speaking raters evaluated 3-minute speech samples recorded by L2 Spanish speakers responding to personally relevant prompts. The raters first used a computer interface which allowed them to increase or decrease the comprehensibility rating as the speech unfolded. The raters then completed a stimulated recall interview, commenting on their thoughts while watching a video capture of their rating. Three distinct rater profiles emerged. Non-dynamic raters (the majority) increased or decreased comprehensibility ratings infrequently. Semi-dynamic raters increased or decreased ratings at a high frequency, but the magnitude of change was small. The two dynamic raters also changed ratings at a high frequency, with a high magnitude of change that was generally in the direction of lower comprehensibility. Most raters reported that their ratings moved towards greater comprehensibility either within the same sample or from one sample to another. Over half of the comments about increasing comprehensibility ratings pertained to discourse. These findings implied that comprehensibility ratings – from the perspective of the listener – are dynamic yet highly variable across raters and that these ratings might ultimately reflect discourse (meaning-making) aspects of interaction.

3. The present study

As discussed previously, comprehensibility ratings provide a good measure of understanding that is sensitive to the linguistic profile of speech; they also offer a useful metric of processing fluency relevant to human judgment. To understand the role of comprehensibility in interactive language use, it would be important to understand whether comprehensibility is a stable phenomenon or whether it fluctuates over time. The raters in Nagle et al.'s (2019) study had completed a one-way listening task, with no possibility to interact with a speaker. However, interactive speech, where interlocutors are reacting to one another in real time, is not only an authentic context of language use but also one that is likely amenable to potential changes in comprehensibility. Therefore, this study's goal was to provide a conversation-centric, time-sensitive view of comprehensibility for both interlocutors in a conversation.

To address this goal, we paired L2 English speakers from different language backgrounds, completing three interactive tasks and rating each other's comprehensibility at approximately 2.5-minute intervals for a total of seven ratings. We examined how the speakers' judgments of each other's comprehensibility changed over time, for each speaker separately and for both conversation partners together in relation to each other's ratings. We also debriefed each speaker to clarify how their interaction and their comprehensibility ratings may have changed over time. Because the raters in Nagle et al.'s (2019) study (although quite variable in their judgments) showed improved ratings within and across the rated speech samples, we expected that comprehensibility ratings would vary across speakers but might show an upward trend as conversation progressed. Based on prior work on speaker–listener adaptation in dialogue (Garrod et al., 2018), we also expected that the two conversation partners might converge on common comprehensibility ratings. Because of the exploratory nature of this study, we made no additional predictions regarding the timing and extent as well as the sources (e.g., task difficulty, language errors) of potential changes in ratings. We asked two broad questions:

1. How do L2 speakers rate each other's comprehensibility over time?
2. Do speakers' ratings of their partners converge or significantly change over time?

4. Method

4.1 Participants

The speakers (M_{age} =25.85 years, *SD*=2.89) included 40 international graduate students (14 women, 26 men) from eight academic disciplines at an English-language university in Canada. The speakers, who reported speaking 17 home languages, had started learning English at a mean age of 8.18 years (*SD*=4.58) and had received all primary and secondary schooling in their home countries. As part of university admission requirements, the speakers took standardized language tests and reported IELTS (31) or TOEFL (9) scores. The TOEFL scores were converted to equivalent IELTS bands using validated conversion metrics (ETS, 2017), with the resulting IELTS scores ranging between 5.5 and 8.0 (*M*=6.84, *SD*=0.62) for the speaking component and between 6.0 and 9.0 (*M*=7.60, *SD*=0.95) for the listening component. As students at a university with a large international enrolment, the speakers indicated that they regularly spoke English (*M*=56.75% daily, *SD*=19.79) and rated themselves as being familiar with accented English (*M*=6.33, *SD*=1.67) on a 9-point scale. Each speaker was randomly paired with a previously unknown partner from a different language background (resulting in 20 pairs), with the constraint that speakers of related languages (e.g., Hindi and Urdu) were paired with partners from other backgrounds (see Appendix A for background information on the speakers' home languages, genders, and ages).

4.2 Tasks

The speakers engaged in three interactive tasks, administered in a fixed order. The first task (3 minutes) was a warm-up task, with the goal of discovering three things the speakers had in common (e.g., a similar hobby), as a way of helping them become familiar with each other. The second task (7 minutes) was a picture story completion task (Galindo Ochoa, 2017). Each speaker had seven scrambled images from a 14-panel picture story. They could not see each other's pictures and had to share their descriptions to produce a common narrative. The story depicted a man who, after winning a large sum of money and purchasing a new house and a car, experienced several unfortunate events, including a car accident and a robbery; the man eventually realized that the money did not make him happy. The final task (7 minutes) was a problem-solving task, where the speakers identified a common set of solutions to challenges experienced by international students. The speakers were encouraged to share their challenges (e.g., long delays in obtaining visas) before proposing common solutions.

4.3 Repeated assessments

During approximately 17 minutes of interaction, the speakers evaluated themselves and their partner for comprehensibility seven times. The speakers also evaluated their own and their partner's communicative anxiety and collaborativeness, but these assessments will not be discussed further. The rating episodes were fairly equally spaced: one at the end of each task (Time 1, 4, and 7) and two additional ratings approximately 2.5 minutes and 5 minutes after the beginning of Task 2 (Time 2 and 3) and after the beginning of Task 3 (Time 5 and 6). The rating scales (100-millimeter lines) were printed next to each rated dimension, one labeled "me" for self-rating and the other labeled "my partner" for the rating of the speaker's partner. The scales contained no markings besides labeled endpoints (*difficult to understand–easy to understand*), and the speakers were asked to mark the point on the line which reflected their judgment. Comprehensibility, introduced to the speakers before the tasks, was defined as a judgment of how much effort it takes to understand what someone is saying. Because this report focuses only on peer-ratings (speakers' evaluations of their partners), self-ratings are not discussed further.

4.4 Procedure

Each pair of speakers was tested individually, and the entire session was audio-recorded. The speakers first completed a background questionnaire. Then, a research assistant (RA) described each rated dimension and explained how to complete the ratings, using practice scales. The RA also advised the speakers that they would complete the scales several times, evaluating the immediately preceding 2–3 minutes of interaction, that they would be stopped periodically during Tasks 2 and 3, and that the ratings were private. The speakers then received the testing booklet, with instructions for each task and seven sets of rating scales printed on separate pages. Each task was introduced immediately before the speakers engaged in it: They first read the printed instructions, then summarized task directions to the RA, and (when applicable) asked clarification questions. The speakers were reminded that the task would stop after the required time had elapsed (3 minutes for Task 1, 7 minutes for Tasks 2 and 3), even if they did not complete their discussion. The RA, who was present in the room during the entire task sequence, stayed away from the speakers during each segment of interaction, using a timer to ensure that task length was comparable across all pairs and that the ratings occurred at evenly spaced intervals. Although the speakers may have felt monitored to some extent, they generally appeared to be focused on completing the tasks. After completing the tasks, both speakers met a different RA in separate rooms and filled out a debrief questionnaire, rating their reactions to the session (100-millimeter scales). Each speaker was then briefly interviewed using guiding questions focusing on their experience during the session.

5. Data analysis

5.1 Coding

The speakers' ratings of each other's comprehensibility were converted to numerical values, defined as the distance (in millimeters) between the left endpoint and the speaker's mark on the scale (out of 100 points). The speakers' rated responses to the debrief questions were also expressed as numeric values (out of 100 points). The recordings of the speakers' interaction were transcribed and then verified by trained RAs to enable a lexical analysis of each speaker's output. Finally, the speakers' interviews were transcribed, with analysis focusing on the speakers' responses to the two most relevant questions, namely, how their interaction changed over time and which aspects of their partners' speech were most difficult to understand. The speakers' comments were coded thematically, following an iterative process, with response categories derived from the content of the transcripts (Gibson & Brown, 2009). The first author initially derived codes for themes and subthemes, then a co-author reviewed the coding and suggested modifications to it, until there was full consensus on all coding decisions.

5.2 Identification of covariates

We identified variables associated with the speakers' comprehensibility ratings so that these variables could be included as covariates in statistical modeling. We first examined the speakers' debrief ratings, on the assumption that the speakers' individual experiences might have impacted their comprehensibility ratings.

The speakers generally found the interaction successful (*M* = 87.5, *SD* = 9.6) and enjoyable (*M* = 91.3, *SD* = 11.4); they considered themselves involved in the tasks (*M* = 92.2, *SD* = 9.3) and found their partners pleasant (*M* = 91.6, *SD* = 10.4); they also expressed interest in speaking to their partners again (*M* = 92.3, *SD* = 12.1) and were satisfied with their performance (*M* = 81.6, *SD* = 13.6). None of the six debrief ratings were associated with comprehensibility (all correlations were below .30), so no debrief category was included in subsequent modeling.

We then targeted the speakers' speaking and listening proficiency, as comprehensibility ratings might reflect each speaker's L2 skill level. Across the 20 pairs, the two paired speakers differed (in absolute values) on average by 0.56 points on the IELTS speaking scale (*SD* = 0.59) and by 1.20 points on the IELTS listening scale (*SD* = 0.70). Although small, these differences could not be regarded as trivial; therefore, both IELTS speaking and listening scores for each speaker were entered as control covariates in subsequent statistical modeling.

We then focused on the speakers' output in each task using lexical profiling (Cobb, 2019) because comprehensibility might reflect each partner's contribution to the dialogue, in terms of its quantity (tokens) and content richness (types). Although token and type frequencies are basic measures of lexical content, they capture substantial amounts of shared variance (38–61%) in listener judgments of L2 speech (Trofimovich & Isaacs, 2012). Because type and token frequencies were highly correlated (*r* = .94), implying their non-independence, only type frequency was used as a covariate, with type values computed separately for each speaker within each segment preceding the rating episode (i.e., before Time 1, between Time 1 and 2, and so on).

In the final check, we examined whether the 20 pairs varied in the amount of time they spent on tasks and in the timing of rating episodes, assuming that comprehensibility might reflect time on task differences. On average, the pairs spent 2 minutes and 46 seconds on Task 1, with some completing this task faster than others (01:04–03:14). However, because few pairs completed Tasks 2 and 3 within the time limit, using (nearly) all of the allotted 7 minutes, their time on task was comparable. The pairs spent on average 7 minutes and 11 seconds on Task 2 (06:58–07:17) and 7 minutes and 8 seconds on Task 3 (06:23–07:17). The repeated ratings also occurred at similar intervals, with the rating episodes spaced about 2.5 minutes apart (02:46–02:37–02:32–02:02–02:35–02:34–02:00). Although time on task was generally consistent across pairs and rating episodes, all statistical models were also re-run with a timing covariate that tracked each pair's deviation from the intended rating time, given that the speakers who rated earlier or later than intended may have evaluated each other differently. Finally, model fit was also evaluated using raw timing (actual time of each speaker's assessment) instead of treating time as an equal-interval variable (seven rating episodes).

5.3 Statistical modeling

The speakers' ratings of each other's comprehensibility were examined through mixed-effects modeling using the lme4 package (Bates et al., 2015) in R version 3.6.1. (R Core Team, 2019). In each set of models, the relevant rating served as the dependent variable, with time (seven rating episodes) as a fixed factor and random intercepts for pairs and for speakers. Model fit was evaluated by performing likelihood ratio tests on pairs of nested models using the ANOVA function, with a more complex model adopted only when it improved fit. For all model parameters, 95% confidence intervals were derived to determine the statistical significance of each parameter (interval does not cross zero). All models included four fixed effects as control covariates: (a) speakers' IELTS speaking score, (b) speakers' IELTS listening score, (c) lexical type frequency in each speaker's output preceding each rating episode, and (d) a speaker-specific time deviation variable capturing whether a rating episode occurred before or after the intended time. All covariates were centered, such that the sample mean was set to 0 and negative values indicated performance below the mean and positive values performance above the mean.

6. Results

6.1 Comprehensibility across time

The first research question asked whether speakers' ratings of their partners' comprehensibility changed over time. Figure 1 illustrates the 40 individual speakers' ratings of their partners' comprehensibility across the seven rating episodes (Time 1–7) with speakers in the same pair shown in the same color. Although different speakers (as rated by their conversation partners) showed varying comprehensibility trajectories, the speakers generally rated each other's comprehensibility high after Task 1 ($M_{\text{Time 1}} = 90.69$, $SD = 11.56$), reduced their ratings during Task 2 ($M_{\text{Time 2}} = 82.14$, $SD = 18.08$, $M_{\text{Time 3}} = 79.78$, $SD = 17.35$), and gradually increased their ratings to approximately the same high initial level by the end of Task 3 ($M_{\text{Time 7}} = 92.31$, $SD = 9.17$). Moreover, Task 2 tended to yield the most variable comprehensibility ratings, with a U-shaped pattern evident for many speakers.

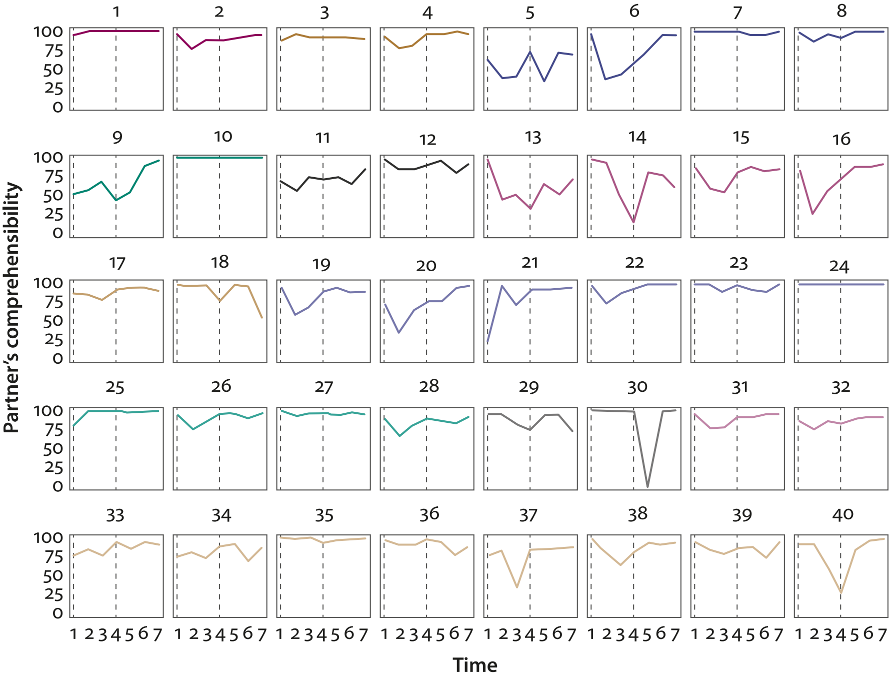


Figure 1. Individual comprehensibility rating trajectories across the seven rating episodes. The vertical dashed lines indicate the three tasks (Task 1: 1, Task 2: 2–4, Task 3: 5–7). Speakers in the same pair are shown in the same color (e.g., 1 and 2, 3 and 4, …, 39 and 40)

To explore the effect of time, we fit four polynomial change models: a null (intercept) model and linear, quadratic, and cubic growth models. Because ratings fluctuated during Task 2, we also fit a piecewise growth model, with time recoded into two dummy variables representing rate of change over Time 1–4 (Tasks 1 and 2) and Time 5–7 (Task 3). In the piecewise model, we estimated linear and quadratic rates of change over Task 2 only, based on the observation that ratings fluctuated most substantially and non-linearly for most participants over that period. This model was equivalent to the cubic growth model in complexity (i.e., had the same number of terms), but the estimated trajectory was slightly different, insofar as the quadratic (U-shaped) function was limited to the first few datapoints. Polynomial time predictors were fit using the poly function to generate orthogonal terms, preventing autocorrelation among linear, quadratic, and cubic slopes.

With respect to the polynomial models, including a higher-order time function significantly improved model fit: null vs linear, $\chi^2(1) = 6.93$, $p = .008$; linear vs. quadratic, $\chi^2(1) = 10.70$, $p = .001$; quadratic vs. cubic, $\chi^2(1) = 7.87$, $p = .005$. Direct comparison of the cubic and piecewise models using likelihood ratio tests was not possible since the models were not nested. We therefore used the Akaike and Bayesian information criteria to select the best-fitting model because these criteria can be used to compare non-nested models fit to the same dataset (Singer & Willett, 2003). The criterion values for the piecewise model were smaller, suggesting that it was a superior fit to the data. By-speaker random slopes were tested for the time terms, but piecewise models including those effects either did not converge or were singular, suggesting overfit. Therefore, only by-speaker and by-pair random intercepts were retained. When we inspected the model residuals, we identified and removed eight datapoints with standardized residuals greater than 2.5 *SD*s (2.86% of the data) and then refit the model. Table 1 summarizes this model, which accounted for 59% of the variance in comprehensibility ratings (marginal $R^2 = .12$, conditional $R^2 = .59$). Figure 2 shows the model-estimated trajectory (dashed line) and observed individual trajectories (solid lines), which display a great amount of variability.

Table 1. Summary of final mixed-effects model for comprehensibility

Fixed effects	Estimate	*SE*	*t*	95% CI	*p*
Intercept	85.64	2.19	39.02	[81.39, 89.89]	<.001
Tasks 1 and 2					
Time linear	−5.11	12.03	−0.42	[−28.57, 18.36]	.67
Time quadratic	52.78	10.77	4.90	[31.61, 73.66]	<.001
Task 3					
Time linear	1.83	0.68	2.68	[0.50, 3.17]	.008
Covariates					
IELTS Speaking	1.80	1.65	1.09	[−1.53, 4.96]	.28
IELTS Listening	1.04	1.51	27.97	[−1.86, 3.99]	.50
Type frequency	0.63	0.75	0.84	[−0.81, 2.12]	.40
Time deviation	−3.25	3.31	−0.98	[−9.66, 3.23]	.33
Random intercepts		***SD***			
Pair		7.80			
Speaker		6.60			

Note. Tasks 1 and 2 linear and quadratic predictors were orthogonal; they should not be interpreted on the original time scale.

As reported in Table 1, the significant coefficient for the orthogonal Task 1 and 2 quadratic slope shows that changes in comprehensibility were not linear over those datapoints, but rather U-shaped, and the significant coefficient for the Task 3 linear slope shows that comprehensibility increased steadily over Task 3. Comprehensibility was independent of the speakers' lexical contribution or their speaking or listening proficiency; these variables did not explain any additional model variance. In addition, the speaker-specific time deviation variable, which captured whether a rating episode occurred before or after the intended time, missed significance. Examining the distribution of model residuals revealed heavy tails.[1]

6.2 Convergence or divergence in comprehensibility

The second research question asked whether the speakers' ratings of each other's comprehensibility became more aligned during interaction. Table 2 summarizes descriptive statistics for comprehensibility ratings at each of the seven rating episodes, separately for the two speakers across the 20 pairs (i.e., for Speaker A vs. Speaker B). The two speakers in each pair were designated as A or B in a random fashion, determined by seat assignment (at opposite sides of a table) upon a speaker's arrival in a testing room.

As shown in Table 2, on average, the two speakers across all pairs appeared the most divergent in each other's comprehensibility ratings during the first rating episode, after Task 1 ($M_{\text{Speaker A}} = 87.32$ vs. $M_{\text{Speaker B}} = 93.90$), such that one speaker in a pair was perceived as being more comprehensible than the other. However, the two speakers generally converged on a common rating approximately 5 minutes into the interaction at Time 2 ($M_{\text{Speaker A}} = 81.20$ vs. $M_{\text{Speaker B}} = 83.19$), and remained fairly aligned after that, except perhaps at Time 5 ($M_{\text{Speaker A}} = 86.03$ vs. $M_{\text{Speaker B}} = 91.89$). Illustrated graphically in Figure 3, the speakers' comprehensibility ratings of their respective partners generally followed the same U-shaped trajectories, but the ratings were substantially mismatched only at the outset of the interaction.

[1] Approximately 19% of the data occurred at the maximum value for comprehensibility (100), suggesting that the dataset was somewhat inflated at the highest range. A zero/one beta regression model was fit to approximate this distribution using the glmmTMB package to account for inflation at either extreme by estimating separate effects for $0 < \text{values} < 1$ and for 1 versus other values. Regression findings confirmed results for the linear model, namely, a significant quadratic trend for Tasks 1–2 ($estimate = 3.07$, $SE = .95$, $z = 3.24$, $p = .001$) and a significant linear trend for Task 3 ($estimate = .17$, $SE = .06$, $z = 2.72$, $p = .007$), except that in this model residuals were normally distributed.

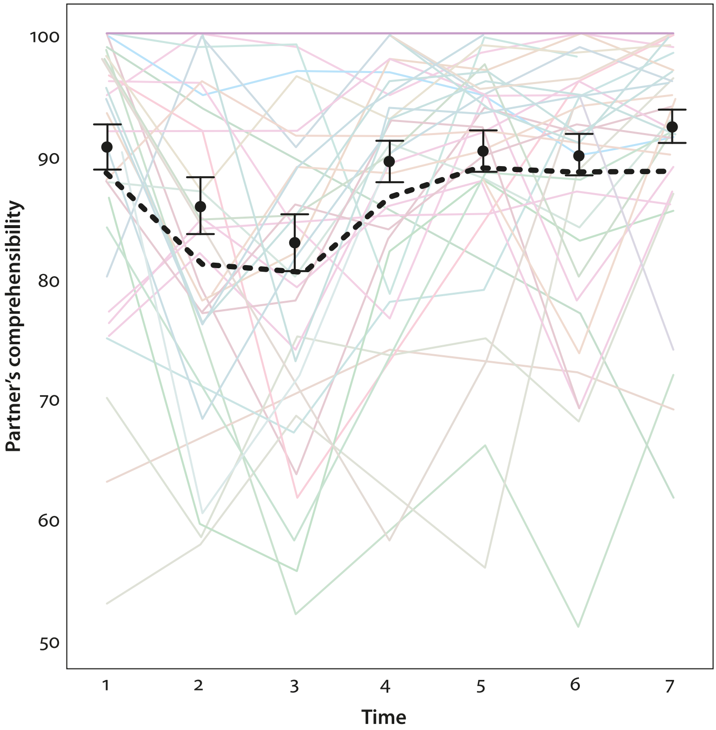


Figure 2. Model-estimated partner comprehensibility trajectory (dashed line) and observed individual trajectories (solid lines). Solid dots designate group mean, and error bars enclose the 95% confidence interval

Table 2. Summary of comprehensibility ratings for Speaker A (as rated by Speaker B) and Speaker B (as rated by Speaker A) across the seven rating episodes

	Speaker A			Speaker B		
Rating	***M***	***SD***	***Range***	***M***	***SD***	***Range***
Time 1	87.32	13.92	53–100	93.90	7.83	75–100
Time 2	81.20	20.50	38–100	83.19	15.47	36–100
Time 3	79.63	17.34	40–100	79.93	17.81	43–100
Time 4	85.18	18.27	32–100	88.56	11.23	58–100
Time 5	86.03	16.58	35–100	91.89	7.95	73–100
Time 6	88.40	12.94	51–100	91.45	9.19	69–100
Time 7	90.88	9.51	69–100	93.82	8.80	62–100

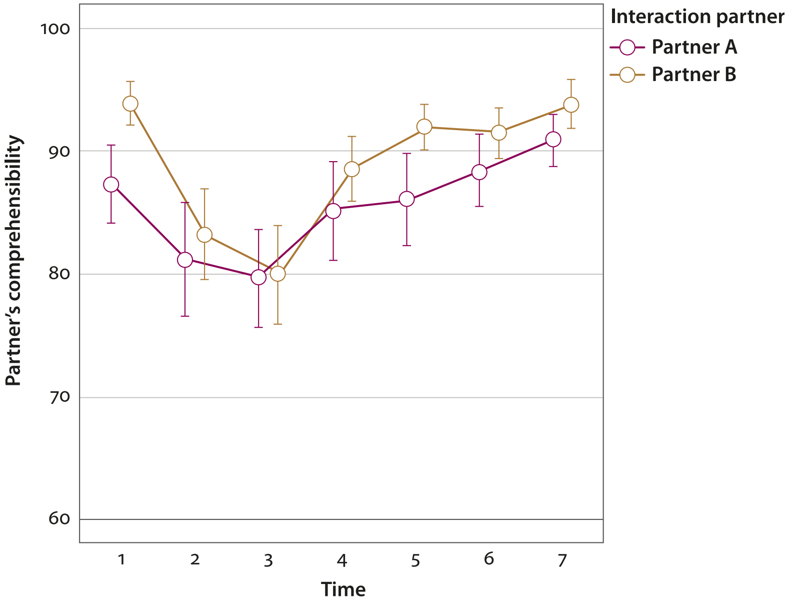


Figure 3. Mean comprehensibility for both speakers in each pair across the seven rating episodes. Vertical bars encompass 95% confidence intervals around the mean values. Speaker A and B designations are random within each pair

Preliminary plotting of group and individual data for within-pairs differences in comprehensibility did not suggest a definitive pattern for change over time; instead, for some pairs, differences in partner comprehensibility diminished over ratings, whereas for others, ratings were most dissimilar near the end of the interaction. Considering this variability, we fit exploratory polynomial models to the absolute value of the within-pair difference in comprehensibility, comparing each model to the baseline (intercept) model. None of these models significantly improved fit over the intercept model. In a follow-up exploratory analysis, which is conceptually similar to the piecewise model reported above, we split the dataset into separate subsets corresponding to Tasks 2 and 3, each with three datapoints, and examined change in alignment in each subset independently. Because the Task 2 and 3 subsets contained only three datapoints, we could only estimate linear and quadratic rates of change.

For Task 2, neither the linear nor quadratic model significantly improved fit over the intercept model. However, for Task 3, the linear model improved fit over the intercept model, albeit marginally, $\chi^2(1) = 4.15$, $p = .04$. Including by-pair random slopes for linear time resulted in singular fit, so the model reported in Table 3 contained only by-pair random intercepts. As before, between-speaker differences in comprehensibility were unrelated to speakers' proficiency and the timing of their ratings. However, lexical characteristics were marginally related to comprehensibility; speakers producing more word types were rated as less comprehensible. Residuals for both final models were normally distributed, with only minor excursions at the upper tail.

Table 3. Summary of fixed effects for between-speaker differences in comprehensibility

Fixed effects	Estimate	*SE*	*t*	95% CI	*p*
Task 2					
Intercept	13.20	2.49	5.31	[8.64, 17.75]	<.001
IELTS Speaking	−0.33	2.84	−0.12	[−5.54, 4.85]	.91
IELTS Listening	2.14	2.07	1.03	[−1.64, 5.92]	.32
Type frequency	−2.44	2.35	−1.04	[−6.99, 1.97]	.30
Time deviation	2.40	4.24	0.57	[−5.35, 10.21]	.58
Task 3					
Intercept	17.70	3.23	5.49	[11.83, 23.89]	<.001
Time linear	−4.45	2.25	−1.97	[−8.91, −0.18]	.06
IELTS Speaking	1.60	2.43	0.66	[−2.79, 6.01]	.52
IELTS Listening	0.58	1.83	0.32	[−2.73, 3.90]	.76
Type frequency	−3.96	1.93	−2.05	[−7.84, −0.45]	.05
Time deviation	5.17	3.77	1.37	[−1.64, 12.00]	.19

6.3 Interview responses

To clarify individual rating patterns, we examined the speakers' interview responses to two questions: how their interaction changed, and which aspects of their partners' speech were most difficult to understand. As shown in Table 4, to explain change over time, the speakers made 58 comments, most of which (44 or 76%) encompassed four categories. In three such categories (33 or 57%), the speakers attributed change to (a) reduced anxiety and increased confidence and comfort, (b) improved or sustained collaboration, or (c) enhanced knowledge of their partner:

- I think maybe at the beginning, we were a bit stressful since we just began and it was like conversation, but then we were more relaxed (S22);
- I think from the first to the last, the collaboration, the sense of collaboration is increased and cooperate more (S9);

- The first activity was about… finding the common things, when you find common things, then it was easier to communicate… so it went easy and easy as time (S1).

The fourth category (11 or 19%) included largely negative comments pertaining to speakers' difficulty with a task, which was exclusively Task 2, or to other methodological issues:

- The second task was a bit difficult to comprehend because of lack of clarity, I wouldn't blame [partner] for [it] because he tried his best to show me the real picture like he was having (S2);
- But with the interruptions, this is something that breaks you… and then you have to rate again, but even with that it's super easy to continue dealing with that (S35).

Table 4. Frequency of comments (*k*) and number of pairs (out of 20) contributing comments

	Change over time			Understanding difficulty		
Coded category	***k***	**%**	**Pairs**	***k***	**%**	**Pairs**
Anxiety, comfort, confidence	14	24.1	11	1	2.3	1
Task effects	11	19.0	10	4	9.3	4
Enhanced knowledge of partner	11	19.0	10			
Increased or sustained collaboration	8	13.8	6			
Accent familiarity	5	8.6	5	6	14.0	6
Shared experience and knowledge	4	6.9	3			
No change, no issue with understanding	3	5.2	3	10	23.3	9
General improvement	2	3.4	2			
Grammar				1	2.3	1
Vocabulary				2	4.7	2
Fluency				2	4.7	2
Voice quality				2	4.7	2
Sufficient explanation and details				4	9.3	4
Pronunciation				11	25.6	10
Total	**58**	**100**		**43**	**100**	

To explain understanding difficulty, the speakers cited pronunciation issues, which made up a quarter (11 or 26%) of the 43 comments produced. Pronunciation issues included generally unclear accent and problems with specific sounds, words, and intonation; vocabulary, grammar, or fluency were rarely mentioned as barriers to understanding, which is unsurprising given that for most speakers the term "accent" encompasses various language issues, including lexical choice, grammatical appropriateness, and issues of fluency or flow:

- I think the accent, his accent was difficult for me (S32);
- Some letters were not pronounced clearly, like… when he was saying "thief" if I heard the "teef" and I have to ask him to repeat it to understand (S6);
- Intonation and pronunciation, I think, she didn't have good intonation (S26).

In another set of comments (11 or 23.3%), the speakers cited no difficulty in understanding each other, largely explained through both partners sharing a cultural background or partners' joint teamwork:

- Because of the community we belong, like it's easy for us to understand… what he is actually going to talk about (S19);
- When he stop speaking, then I speak; sometimes when I stop speaking, he speak… we cooperate well (S25).

Accounting for 9–14% of the comments, other reasons for understanding difficulty included task-specific issues (again limited to Task 2), familiarity with partners' accent, and partners' ability to express ideas clearly or provide sufficient detail:

- Accent a little bit different, but I get used to this… it's like I understand it's not perfect British English that I learned at school (S22);
- He's not explaining the part of the picture… he's giving only one two three pictures scenarios (S31).

More importantly, the speakers' individual comprehensibility ratings (plotted in Figure 1) did not appear to unambiguously map onto the stated reasons for how their communication changed or which issues contributed to difficulty in understanding. For example, of the seven speakers rated consistently as being highly comprehensible (1, 3, 7, 10, 24, 27, and 35 in Figure 1), there were only two cases where the partner cited no problem with understanding the speaker. Similarly, across the speakers whose comprehensibility was rated as changeable (dynamic trajectories in Figure 1), 11 partners reported no problems contributing to difficulty in understanding these speakers or reported no change to communication over time.

7. Discussion

This exploratory study's goal was to examine whether comprehensibility could be viewed as a dynamic, time-sensitive construct for both interlocutors in a conversation and to explore whether comprehensibility ratings might be co-dependent on both interlocutors and thus subject to convergence or divergence effects over time. We found evidence for a dynamic change in comprehensibility consistent with an exponential (U-shaped) trendline which was independent of speakers' proficiency, lexical contribution, or the timing of a rating episode. Although the speakers' comprehensibility judgments displayed a great amount of inter-individual variability, they rated their partners' comprehensibility as generally high after Task 1, their ratings then dropped during Task 2 and increased gradually throughout Task 3. In terms of the relationship between interlocutors' ratings, although the best-fitting model showed no significant time effect, the absolute differences in mutual ratings seemed to diminish over time and tasks, approximating a linear function, especially during Task 3, suggesting that the speakers' ratings showed more similarity over time. Based on interview comments, the most frequent changes to communication patterns were decreased anxiety and increased confidence, improved collaboration, and enhanced knowledge of the partner. The most cited issues leading to difficulties in understanding were various pronunciation issues, task-specific influences, and partner's ability to provide sufficient content detail.

7.1 Time- and task-sensitive view of comprehensibility

In their study exploring how raters' comprehensibility ratings evolved over time, Nagle et al. (2019) provided a micro perspective on comprehensibility as a time-sensitive construct, arguing that comprehensibility judgments displayed several properties of dynamic systems (de Bot et al., 2007), including change over time and nonlinearity. For instance, comprehensibility judgments in that study were variable both within and across the raters and displayed nonuniformity, in the sense that different types of linguistic issues (e.g., phonemic errors, lexical substitutions), with their particular timing and location in the evolving narrative, elicited different reactions from different raters. To complement this micro-level view of comprehensibility, the present findings offer a global, macro-level perspective, demonstrating that comprehensibility ratings for both speakers in a conversation, while overall highly variable, seem to fluctuate in tandem in extended communication. The two macro variables emerging from this dataset with relevance to comprehensibility are time and task.

That comprehensibility ratings are sensitive to time (understood broadly as listeners' cumulative experience with a speaker's speech) is unsurprising. To begin with, time might have a negative influence on comprehensibility, so that speech evokes more effortful processing for the listener, at least early in a conversation. For example, raters sometimes assign harsher ratings when they evaluate the same sentence-length speech samples again, because raters might become increasingly aware of how the speakers' output differs from the language expected by the rater (Flege & Fletcher, 1992). Similarly, there seems to be little consistency between raters' evaluations of separate short sentences by the same speaker, suggesting that ratings of shorter speech samples might not be representative of ratings of longer discourse produced by the same speaker (Munro, 2018).

Time might also impact comprehensibility positively, such that, as interaction proceeds, speech becomes less effortful for the listener. For instance, raters with greater linguistic exposure or experience (language teachers, multilinguals) generally assign higher ratings than those with less exposure or experience (Kang, 2012; Saito & Shintani, 2016). An upward trend in comprehensibility would also be compatible with the notion that listeners' perceptual categories are highly adaptive to recent experience (Baese-Berk, 2018). In the end, both negative and positive time-bound forces may have been at play here, yielding a U-shaped comprehensibility function, with negative influences operating early in the interaction, until a certain temporal threshold was reached, and positive influences acting as comprehensibility attractors later on.

Comprehensibility ratings also seemed to depend on the communicative task performed by interlocutors, on the assumption that different tasks impose greater or lesser demands on the speaker and thus increase or decrease processing effort for the listener. Increased task difficulty likely elicits more sophisticated language from speakers, while also increasing opportunities for them to make errors or experience a communication breakdown (Robinson, 2005), which may explain why raters experience greater processing effort in evaluating more complex tasks (Crowther et al., 2015b). In this study, the dip in comprehensibility following the first rating (illustrated by a quadratic time function for Tasks 1–2) was likely due to higher cognitive demands in the second task, with ratings continuing to rise as speakers moved through an easier task (shown by a positive linear time function for Task 3).

In terms of task difficulty, the initial task had low cognitive demands because speakers had an unlimited range of possible commonalities to consider. The second task was more complex due to the need to exchange nonshared information by identifying and describing referents in 14 scrambled images (a conclusion also supported in interview comments). The final task was less demanding because partners had shared access to the initial information and could complete the task by co-constructing agreed-upon solutions. Until the effects of time and task are disentangled in future work by rotating task order across speakers, our interim conclusion is that comprehensibility trajectories reflect individual and joint contributions of interlocutors becoming accustomed to each other and to specific tasks and their features over time. A key qualification here is that such interlocutor experiences across tasks and time appear to be subject to vast amounts of interindividual variability, which also needs to be explained in future work.

7.2 Between-speaker alignment in comprehensibility

Comprehensibility ratings for the two speakers engaged in interaction appeared to be co-dependent, showing a trend for convergence over time. Between-speaker comprehensibility differences were approximately 15–20 points on a 100-point scale and were highly variable. However, these differences generally decreased over time, particularly during Task 3, dropping below a 10-point difference by the end of the interaction. This novel finding extends prior work on interactive alignment (Garrod et al., 2018) to include interlocutor alignment in comprehensibility. Interactive alignment reflects a phenomenon whereby interlocutors converge on common speech patterns, driven by such social forces as accommodation to an interlocutor (Giles & Ogay, 2007) and psychological mechanisms of priming (Garrod et al., 2018), with alignment involving multiple features of speech, including utterance length, speech rate, phonetic realizations of segments and words, volume, and pausing frequencies and lengths (Garrod et al., 2018). Convergence in various speech patterns has also been attested among L2 speakers and has been shown to depend on speech style and speaker proficiency (Berry & Ernestus, 2018). The finding that frequency of word types was negatively associated with speaker convergence in comprehensibility highlights another variable that might modulate alignment, in this case, by increasing interlocutors' effort in understanding speech with increased lexical content.

The obtained evidence for speaker convergence in comprehensibility is also consistent with prior research on listener adaptation to foreign accent (Baese-Berk, 2018). For instance, listeners rapidly get attuned to the speech of unfamiliar L2 speakers, often requiring just over a minute of experience (Clarke & Garrett, 2004). Xie et al. (2018) recently extended these findings to show that listeners improve quickly (in a matter of minutes) in speed and accuracy of comprehension of unfamiliar L2 speakers, arguing that long- and short-term adaptations to L2 speech might be driven by similar mechanisms. Our finding of a rapid convergence in interlocutors' comprehensibility ratings, which generally occurred within 1–3 minutes of their experience in the initial task (see Figure 3), is suggestive of a parallel phenomenon for comprehensibility. Adaptation to L2 comprehensibility (at least for L2 speakers) might involve interlocutors engaging in a process of adjusting their expectations of the effort involved in understanding their partners, by checking these expectations against the actual linguistic evidence available in discourse (for a potential model, see Kleinschmidt & Jaeger, 2015). Because the discourse in dialogue is usually co-constructed by both interlocutors (e.g., through turn-taking and feedback as part of attaining a common interactive goal) and likely involves interlocutors predicting upcoming content and potential language errors, it is reasonable that L2 interlocutors (especially those at comparable L2 skill levels) would arrive at a shared, conversation-specific (rather than speaker-specific) view of comprehensibility. As shown in Figure 3, once a shared understanding of comprehensibility has been reached (which might require more time for some pairs than for others), this shared rating of comprehensibility is what describes both partners' conversational experience across time and task.

8. Limitations and future work

A major limitation of this work, which prevented us from making specific predictions beyond asking exploratory questions, is that tasks were not rotated across speakers. As discussed previously, it is important to examine whether similar U-shaped comprehensibility trajectories would emerge when speakers engage in communicative tasks ordered differently, clarifying how interlocutors' cumulative shared experience impacts their comprehensibility ratings in tasks that increase versus decrease in cognitive difficulty across time. Similarly, the speakers' comments regarding changes in their interaction patterns and reasons for difficulty in understanding their partners did not unambiguously explain the speakers' comprehensibility rating trends. For instance, the speakers may not have been aware of how and why their perceived effort of understanding their partners varied, which would implicate an implicit component to ratings. More likely, however, the speakers did not possess the needed terminology to describe their thought processes and largely resorted to the categories made salient to them (see Table 4), through either the experimental procedure (anxiety, collaboration) or conversation tasks (understanding, getting to know partners). The link between interaction-based comprehensibility ratings and interlocutor awareness of comprehensibility should be investigated further, using different combinations of interlocutors that vary in language proficiency and experience.

With respect to interactive alignment, as suggested by an anonymous reviewer, between-speaker convergence or divergence in comprehensibility could be potentially misleading, in the sense that speakers may have given the impression of convergence or divergence because they approached the rating task using different criteria, showed varying degrees of rating severity, or tended to avoid extreme rating values (thus demonstrating a regression-to-the-mean effect). Future work should therefore revisit the validity of interactive ratings of comprehensibility by ensuring (at minimum) that raters are trained on the use of the rating scale to the point of calibrated performance. It would also be interesting to explore potential effects of cognitive workload on alignment in comprehensibility ratings. For example, certain interactive tasks might be particularly prone to highlighting partner-specific comprehensibility issues in interaction, preventing or delaying convergence. Similarly, it might be useful to explore long-term effects of interlocutors' extended conversational experience on their perception of comprehensibility, focusing on speakers' judgments of the same and new partners in another instance of interaction, after a delay. In light of the alignment between both partners' comprehensibility scores in extended interaction, it might also be fruitful to examine the validity of a joint (rather than speaker-specific) measure of comprehensibility for both partners in a conversational dyad. Finally, comprehensibility ratings, as useful measures of listener understanding and listener processing fluency, could be examined in relation to such conversation phenomena as speakers' engagement in dialogue, their participation patterns, or their affective response to the task or their partner, to clarify the role of processing effort in interlocutor experience in dialogue.

9. Conclusion

Over the last 25 years, comprehensibility ratings have become a valuable metric that captures various facets of listeners' experience with L2 speech, implicated in multiple social, linguistic, and psychological phenomena. Our goal was to extend prior work on comprehensibility by providing a conversation-centric, dynamic view of this construct in interaction. Our findings imply that listeners' judgements of L2 comprehensibility can change in real time according to listeners' immediate experience, particularly for listeners in interactive speaking tasks, and that such judgments may be subject to convergence effects over time. These initial results call for rigorous future research in order to understand whether comprehensibility – as a proxy for listeners' effort in understanding speech – could capture many other important real-life dimensions of L2 speakers' performance (communication anxiety collaborativeness, engagement, affective response) as they evolve in interaction in real time.

Acknowledgements and open materials and data statement

We are grateful to Clinton Hendry for help with data collection, Aki Tsunemoto for help with data analyses, and to the anonymous reviewers and the journal editor for the insightful comments and suggestions that helped us refine this article. All materials and data from this study are publicly accessible through the Open Science Framework at https://osf.io/bpzgm and the IRIS digital repository at http://www.iris-database.org.

References

Baese-Berk, M. (2018). Perceptual learning for native and non-native speech. In K. D. Federmeier & D. G. Watson (Eds.), *The psychology of learning and motivation: Current topics in language* (pp. 1–29). Academic Press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67, 1–48. https://doi.org/10.18637/jss.v067.i01

Bergeron, A., & Trofimovich, P. (2017). Linguistic dimensions of accentenedness and comprehensibility: Exploring task and listener effects in second language French. *Foreign Language Annals*, 50, 547–566. https://doi.org/10.1111/flan.12285

Berry, G. M., & Ernestus, M. (2018). Phonetic alignment in English as a lingua franca: Coming together while splitting apart. *Second Language Research*, 34, 343–370. https://doi.org/10.1177/0267658317737348

Brennan, S. E., Kuhlen, A. K., & Charoy, J. (2018). Discourse and dialogue. In S. L. Thompson-Schill (Ed.), *The Stevens' handbook of experimental psychology and cognitive neuroscience* (pp. 145–209). Wiley. https://doi.org/10.1002/9781119170174.epcn305

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English, *Journal of the Acoustical Society of America*, 116, 3647–3658. https://doi.org/10.1121/1.1815131

Cobb, T. (2019). VocabProfilers [computer program]. https://www.lextutor.ca/vp

Crowther, D., Trofimovich, P., Isaacs, T., & Saito, K. (2015b). Does speaking task affect second language comprehensibility? *The Modern Language Journal*, 99, 80–95. https://doi.org/10.1111/modl.12185

Crowther, D., Trofimovich, P., Isaacs, T., & Saito, K. (2018). Linguistic dimensions of L2 accentedness and comprehensibility vary across speaking tasks. *Studies in Second Language Acquisition*, 40, 443–457. https://doi.org/10.1017/S027226311700016X

Crowther, D., Trofimovich, P., Saito, K., & Isaacs, T. (2015a). Second language comprehensibility revisited: Investigating the effects of learner background. *TESOL Quarterly*, 49, 814–837. https://doi.org/10.1002/tesq.203

de Bot, K., Lowie, W., & Verspoor, M. (2007). A Dynamic Systems Theory approach to second language acquisition. *Bilingualism: Language and Cognition*, 10, 7–21. https://doi.org/10.1017/S1366728906002732

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, 19, 1–16. https://doi.org/10.1017/S0272263197001010

Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. John Benjamins. https://doi.org/10.1075/lllt.42

Dragojevic, M., Giles, H., Beck, A.-C., & Tatum, N. T. (2017). The fluency principle: Why foreign accent strength negatively biases language attitudes. *Communication Monographs*, 84, 385–405. https://doi.org/10.1080/03637751.2017.1322213

ETS. (2017). *TOEFL iBT® and IELTS® academic module scores: Score comparison tool.* http://www.ets.org/toefl/institutions/scores/compare

Flege, J., & Fletcher, K. (1992). Talker and listener effects on the perception of degree of foreign accent. *Journal of the Acoustical Society of America*, 91, 370–389. https://doi.org/10.1121/1.402780

Galindo Ochoa, J. A. (2017). The effect of task repetition on Colombian EFL students' accuracy and fluency (Unpublished master's thesis). Concordia University, Montreal.

Garrod, S., Tosi, A., & Pickering, M. J. (2018). Alignment during interaction. In. S.-A. Rueschemeyer & M. G. Gaskell (Eds.), *The Oxford handbook of psycholinguistics* (pp. 575–593). Oxford University Press.

Gibson, W., & Brown, A. (2009). *Working with qualitative data.* Sage. https://doi.org/10.4135/9780857029041

Giles, H., & Ogay, T. (2007). Communication Accommodation Theory. In B. B. Whaley & W. Santer (eds.), *Explaining communication: Contemporary theories and exemplars* (pp. 293–309). Lawrence Erlbaum.

Graf, L. K. M., Mayer, S., & Landwehr, J. R. (2018). Measuring processing fluency: One versus five items. *Journal of Consumer Psychology*, 28, 393–411. https://doi.org/10.1002/jcpy.1021

Kang, O. (2012). Impact of rater characteristics and prosodic features of speaker accentedness on ratings of international teaching assistants' oral performance. *Language Assessment Quarterly*, 9, 249–269. https://doi.org/10.1080/15434303.2011.642631

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, 122, 148–203. https://doi.org/10.1037/a0038695

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition, and Neuroscience*, 31, 32–59. https://doi.org/10.1080/23273798.2015.1102299

Matsuura, H., Chiba, R., & Fujieda, M. (1999). Intelligibility and comprehensibility of American and Irish Englishes in Japan. *World Englishes*, 18, 49–62. https://doi.org/10.1111/1467-971X.00121

Munro, M. J. (2018). Dimensions of pronunciation. In O. Kang, R. I. Thomson, & J. M. Murphy (Eds.), *The Routledge handbook of contemporary English pronunciation* (pp. 413–431). Routledge.

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, 45, 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Nagle, C., Trofimovich, P., & Bergeron, A. (2019). Toward a dynamic view of second language comprehensibility. *Studies in Second Language Acquisition*, 41, 647–672. https://doi.org/10.1017/S0272263119000044

Newman, E. J., Sanson, M., Miller, E. K., Quigley-McBride, A., Foster, J. L., Bernstein, D. M., & Garry, M. (2014). People with easier to pronounce names promote truthiness of claims. *PLoS ONE*, 9(2), e88671. https://doi.org/10.1371/journal.pone.0088671

O'Brien, M. G. (2014). L2 learners' assessments of accentedness, fluency, and comprehensibility of native and nonnative German speech. *Language Learning*, 64, 715–748. https://doi.org/10.1111/lang.12082

Pakhomov, S.V., Kaiser, E.A., Boley, D.L., Marino, S.E., Knopman, D.S., & Birnbaum, A.K. (2011). Effects of age and dementia on temporal cycles in spontaneous speech fluency. *Journal of Neurolinguistics*, 24, 619–635. https://doi.org/10.1016/j.jneuroling.2011.06.002

R Core Team. (2019). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org

Reber, R., & Greifeneder, R. (2017). Processing fluency in education: How metacognitive feelings shape learning, belief formation, and affect. *Educational Psychologist*, 52, 84–103. https://doi.org/10.1080/00461520.2016.1258173

Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *International Review of Applied Linguistics in Language Teaching*, 43, 1–32. https://doi.org/10.1515/iral.2005.43.1.1

Saito, K., & Shintani, N. (2016). Do native speakers of North American and Singapore English differentially perceive comprehensibility in second language speech? *TESOL Quarterly*, 50, 421–446. https://doi.org/10.1002/tesq.234

Saito, K., Trofimovich, P., & Isaacs, T. (2017). Using listener judgements to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics*, 38, 439–462. https://doi.org/10.1093/applin/amv047

Sanchez, C.A., & Jaeger, A.J. (2015). If it's hard to read, it changes how long you do it: Reading time as an explanation for perceptual fluency effects on judgment. *Psychonomic Bulletin and Review*, 22, 206–211. https://doi.org/10.3758/s13423-014-0658-6

Sanchez, C.A., & Khan, S. (2016). Instructor accents in online education and their effect on learning and attitudes. *Journal of Computer Assisted Learning*, 32, 494–502. https://doi.org/10.1111/jcal.12149

Sheppard, B.E., Elliott, N.C., & Baese-Berk, M.M. (2017). Comprehensibility and intelligibility of international student speech: Comparing perceptions of university EAP instructors and content faculty. *Journal of English for Academic Purposes*, 26, 42–51. https://doi.org/10.1016/j.jeap.2017.01.006

Singer, J.D., & Willett, J.B. (2003). *Applied longitudinal data analysis*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195152968.001.0001

Song, H., & Schwarz, N. (2008). If it's hard to read, it's hard to do: Processing fluency affects effort prediction and motivation. *Psychological Science*, 19, 986–988. https://doi.org/10.1111/j.1467-9280.2008.02189.x

Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition*, 15, 905–916. https://doi.org/10.1017/S1366728912000168

Xie, X., Weatherholtz, K., Bainton, L., Rowe, E., Burchill, Z., Liu, L., & Jaeger, T.F. (2018). Rapid adaptation to foreign-accented speech and its transfer to an unfamiliar talker. *Journal of the Acoustical Society of America*, 143, 2013–2031. https://doi.org/10.1121/1.5027410

Appendix A. Background information for speaker pairs

	Speaker A			Speaker B		
Pair	Native language	Gender	Age	Native language	Gender	Age
1	Farsi	male	26	Tamil	male	24
2	Hindi	female	24	Malayalam	male	25
3	Vietnamese	male	31	Arabic	female	25
4	Mandarin	male	24	Farsi	female	26
5	Farsi	male	30	Bengali	male	27
6	Hindi	female	24	Mandarin	female	23
7	Kannada	male	25	Portuguese	male	24
8	Gujarati	female	27	Azeri	male	25
9	Arabic	male	26	Punjabi	female	24
10	Tamil	male	24	Hindi	male	23
11	Hindi	male	23	Russian	female	28
12	Hindi	female	24	Farsi	male	28
13	Mandarin	female	24	Farsi	male	24
14	Nepali	male	23	Tamil	male	22
15	Farsi	male	27	Hindi	female	27
16	Hindi	male	26	Farsi	male	35
17	Tulu	female	25	Farsi	male	29
18	Portuguese	male	32	Farsi	male	30
19	Mandarin	female	23	Bengali	male	29
20	Urdu	male	22	Kannada	female	26

International intelligibility revisited[*]

L2 realizations of NURSE and TRAP and functional load

Veronika Thir
University of Vienna

The Lingua Franca Core (LFC) proposes that NURSE is the only vowel quality important for international intelligibility, yet research findings regarding this issue are mixed. Moreover, it is unclear whether phonetic (rather than phonemic) substitutions of NURSE also affect international intelligibility more negatively than other phonemic vowel substitutions, though this seems unlikely on the basis of considerations of functional load (FL). This study compares the international intelligibility of two vowel substitutions typical of Austrian learners of English: the phonetic replacement of NURSE with a rounded and diphthongized vowel, and the phonemic replacement of TRAP with a vowel close to cardinal [ɛ]. The findings suggest that, contrary to the LFC but in line with FL considerations, the phonetic substitution of NURSE is more intelligible to an international audience than the substitution of TRAP with [ɛ]. However, differences in intelligibility between the two substitutions were largely 'neutralized' once contextual support was available.

Keywords: English as a lingua franca (ELF), international intelligibility, Lingua Franca Core (LFC), functional load (FL), TRAP, NURSE

1. Introduction

Most verbal exchanges in English take place between native speakers (NSs) and nonnative speakers (NNSs) from different lingua-cultural backgrounds. The significant phonetic-phonological heterogeneity involved in such English as a lingua franca (ELF) exchanges often raises concerns for mutual intelligibility. However,

[*] An earlier version of this article was published as part of a special issue of the *Journal of Second Language Pronunciation* 6(3). https://doi.org/10.1075/jslp.20012.thi

https://doi.org/10.1075/bct.121.09thi

intelligibility, i.e., "the extent to which a speaker's message is actually understood by a listener" (Munro & Derwing, 1995, p. 76), is not necessarily impeded by a foreign accent (Munro & Derwing, 1995; Derwing & Munro, 1997), which highlights the need for research identifying those L2 pronunciation features that interfere with learners' intelligibility. With regard to *international intelligibility* (i.e., intelligibility in ELF contexts), Jenkins' (2000) proposed the 'Lingua Franca Core' (LFC), a set of English pronunciation features she argues are crucial for maintaining intelligibility among ELF users. This core includes most English consonant sounds, preservation of word-initial consonant clusters, aspiration of plosives, nuclear stress, chunking, and vowel length contrasts. The only vowel quality included in the core is the NURSE vowel /ɜː/, found in non-rhotic Standard British pronunciation (e.g., *nurse*).

Jenkins' study was qualitative and exploratory; she therefore conceded that the LFC required further empirical consolidation (2000). However, subsequent research exploring the LFC's claim that vowel quality (apart from /ɜː/) was irrelevant for international intelligibility proved largely inconclusive. Deterding and Kirkpatrick's study (2006) and a small-scale investigation by Luchini and Kennedy (2013) supported the importance of /ɜː/ for international intelligibility over other English vowels. However, in his examination of ELF interactions among pre-dominantly Asian users of English, Deterding (2013) found vowel quality played a minor role in comprehension difficulties, which also applied to variations in the production of /ɜː/. Similar results were obtained by Zoghbor (2010) who examined the international intelligibility of Arab speakers of English; Cole (2002) also found a limited effect for vowel quality in a small-scale study.

Some researchers argue for the importance of vowel quality for international intelligibility. Kennedy (2012) identified it as the primary source of unintelligibility in ELF interactions. Notably, all examples of problematic vowel productions in her data involved vowels other than /ɜː/, which suggests that /ɜː/ did not occupy a more important place regarding international intelligibility than other vowels in the exchanges she examined. Moreover, small-scale studies by O'Neal (2015) and Kim and Billington (2018) provide examples regarding the (potential) significance of vowel qualities other than /ɜː/ for intelligibility among ELF speakers.

The inconclusive research findings regarding the role of vowel quality in general and of /ɜː/ in particular for international intelligibility are due to several factors. First, most studies mentioned were qualitative examinations of interactive speech data involving a relatively small number of participants and/or tokens of unintelligibility. Thus, while Deterding's (2013) study analyzed 183 tokens of unintelligibility, 138 of which were pronunciation-related, it involved only nine speaker-listeners. The data analyzed by Deterding and Kirkpatrick (2006) came from 20 participants, yet their claims were based on only five pronunciation-related tokens of unintelligibility, one of which involved /ɜː/. Kennedy's (2012) comparatively larger study analyzed interactions of 20 NNS dyads using ELF with the help of stimulated recall. Out of 161 participant comments regarding instances of unintelligibility in the interactional data, 54 were pronunciation-related, yet participants were unable to specify the exact cause of the problem in almost half of them, which makes it hard to draw firm conclusions. Finally, Cole (2002), Luchini and Kennedy (2013), and O'Neal (2015) did not systematically quantify instances of unintelligibility in their data but instead analyzed a small number of selected or 'illustrative' episodes of unintelligibility, and Kim and Billington (2018) discussed a single instance of miscommunication (albeit a crucial one with potentially serious consequences). Though all these studies have contributed to research on intelligibility in ELF, for example, by highlighting how ELF users negotiate meaning interactively, they do not constitute a firm basis for generalizations regarding the international intelligibility of certain pronunciation features. That is, while they have been important in generating hypotheses regarding international intelligibility such as the LFC proposal, these hypotheses are yet to be tested on the basis of larger amounts of data, especially if they are to be translated into pedagogic practice (see also Derwing & Munro, 2015).

The second factor giving rise to inconsistent findings are methodological limitations of the research summarized above, mostly based on what Sewell (2017) terms the 'a posteriori approach', in which pronunciation features crucial to maintaining international intelligibility are identified through observations of communication problems in interactional data (Deterding, 2013; Deterding & Kirkpatrick, 2006; Jenkins, 2000; Kim & Billington, 2018; Luchini & Kennedy, 2013; O'Neal, 2015). Sewell highlights two problems; the first is the 'co-occurrence problem', i.e., the difficulty in determining the exact sources of unintelligibility due to "multiple interacting factors at work" (p. 61). This is evident in Deterding's (2013) data analysis, where tokens of unintelligibility often involve variation in more than one pronunciation feature and are occasionally related to problems at more than one linguistic level. However, as Sewell (2017) notes, this is a necessary consequence of working with natural speech data, the pay-off being higher ecological validity. He argues that the merits of this approach should not be overlooked, but that it should be complemented by findings obtained under more controlled conditions, to identify cause-effect relationships in intelligibility problems more clearly.

Another drawback of the *a posteriori* approach addressed by Sewell (2017) relates to the irrelevance of certain pronunciation features in international contexts. The absence of instances of unintelligibility caused by a particular pronunciation feature is commonly viewed as evidence of its unimportance for international intelligibility, as has been argued with regard to vowel quality. According to Sewell, this conclusion seems apt only if vowel modifications occur frequently in the data, for their rare occurrence may point to their importance for intelligibility, with speakers having acquired vowel phonemes for their interlocutor's receptive needs. In addition, few modifications will likely lead to a smaller number of communication breakdowns – unless a feature should, for some reason, be particularly crucial to maintaining intelligibility. Thus, "absence of evidence is not evidence of absence" (Sewell, 2017, p. 62), that is, the fact that intelligibility problems are rarely associated with vowel quality modifications does not entail their insignificance for international intelligibility. To establish the importance of a particular pronunciation feature in international communication, it is necessary to compare its frequency in a dataset with the number of instances of unintelligibility it provoked. Clearly, it makes a difference whether a feature causes unintelligibility in 10%, 50%, or 90% of all instances in which it occurred. Such percentages allow for a more meaningful comparison of the impact of pronunciation features on intelligibility, but this type of quantification may not be feasible when analyzing several hours of interactive speech data. It can, however, be easily completed in more controlled, experimental approaches, such as the one presented here.

Another unresolved issue is whether all substitutions of /ɜː/ are equally problematic (if at all) for international intelligibility. This question relates to the concept of functional load (FL), a measure of the importance of sounds and phonemic contrasts for intelligibility in a language. The term most commonly refers to the number of minimal pairs (MPs) that exist for a certain phonemic contrast (e.g., Catford, 1987), though more sophisticated approaches to calculating FL exist as well (see Brown, 1988). In studies that reported variations in /ɜː/ to be detrimental to intelligibility among ELF users, the sound was mostly substituted by an (approximation to an) open vowel phoneme: [ɑː] in Jenkins (2000) and Deterding & Kirkpatrick (2006); [a] in Luchini & Kennedy (2013); and "an open vowel" (p. 65) (presumably [ɑ] or [a], since the phrase 'early morning' was misidentified as 'alimony') and [ɪə] in Deterding (2013). However, the replacement of /ɜː/ by /eɪ/ did not seem to cause intelligibility problems in Zoghbor (2010). These findings can be partially explained on the basis of FL: while /ɜː/-/ɑː/ and /ɜː/-/ʌ/ have an intermediate FL (Catford, 1987; Brown, 1988), the contrast /ɜː/-/eɪ/ (which seems to only rarely cause problems for international intelligibility) has a markedly lower FL (Catford, 1987), and /ɜː/-/eɪ/ is not even included in Catford's (1987) and Brown's (1988) FL scales. Thus, FL may provide a useful theoretical basis for the findings of ELF intelligibility studies (see also Sewell 2017). Moreover, the recommendations of the LFC regarding /ɜː/ should be tested in light of the sound substitutions that particular learners make, since not every phonemic substitution of /ɜː/ may cause intelligibility problems in ELF contexts. In the absence of further empirical evidence, however, teachers who wish to follow the LFC proposal are sometimes generally recommended to teach /ɜː/, even if the substitution their learners use is arguably *phonetic* rather than phonemic (e.g., Berger, 2010). From a FL perspective, it seems unlikely that phonetic substitutions of /ɜː/, whose FL is effectively zero, would often cause problems in ELF contexts, or that they would do so more frequently than *phonemic* replacements of other English vowel sounds. Since this hypothesis has important implications for teaching practice, it must be tested empirically.

The present study thus compares the international intelligibility of a phonetic substitution of /ɜː/ common for Austrian learners of English to that of a phonemic vowel substitution common for this group. Learners with Austrian German as their L1 tend to replace /ɜː/ with a rounded vowel and often additionally diphthongize it, resulting in the vowel [øə] (Richter, 2019), which does not create a phonemic merger. By contrast, their tendency to raise /æ/ to a position close to cardinal [ɛ] leads to the loss of the phonemic distinction in MPs such as *bad-bed*. Comparing the international intelligibility of these vocalic substitutions is of interest for two reasons. First, they are by no means limited to Austrian speakers of English: similar realizations of /ɜː/ can be found for example amongst Cantonese (Chan & Li, 2000), Dutch (Collins & Vandenbergen, 2000) and German learners more generally (O'Connor, 1980), and the raising of /æ/ along with the merger of /æ/ and /ɛ/ also occurs amongst users of English from different Outer and Expanding Circle countries (see e.g., Chan & Li, 2000; Komar, 2017). Second, such a comparison is interesting in terms of FL considerations: whereas the contrast /æ/-/ɛ/ has a comparatively high FL (the highest on Brown's (1988) scale and an intermediate FL on Catford's scale (1987)), the FL of the contrast /ɜː/-[øə] is effectively zero, since it does not serve to distinguish meaning in English. One would therefore expect the replacement /æ/ → [ɛ] to more frequently cause intelligibility problems in ELF communication than /ɜː/ → [øə]. Contrasting these two substitutions may lead to further insights into the explanatory potential of FL for international intelligibility. This paper therefore addresses the following overarching research question:

(RQ1) Does the Austrian replacement of /ɜː/ with [øə], in line with the LFC but contrary to FL considerations, inhibit international intelligibility more than the substitution of /æ/ with [ɛ]

In connection with this question, it is important to recognize that not every word in a language forms part of an MP. Thus, the substitution /æ/ → /ɛ/ may have a different effect on intelligibility according to whether it results in an existing English word, as for example in *bad* → *bed* (MP words), or whether it does not, as in *flat* → **flet* (non-MP words). The second research question therefore asks:

(RQ2) Are words involving the substitution of /ɜː/ with [øə] less intelligible to an international audience than *both* MP and non-MP words involving the substitution /æ/ → /ɛ/?

The effect of MP-status on intelligibility is of particular interest since MPs are central to the concept of FL and moreover are frequently prioritized in pronunciation textbooks due to the widespread belief that L2 learners run a higher risk of being misunderstood when using phonemic substitutions in MP words than in non-MP words (Levis & Cortes, 2008). On the basis of this assumption, MP words involving the substitution /æ/ → /ɛ/ should be more difficult to understand than non-MP words involving the same substitution, and non-MP words involving a phonetic sound substitution, such as /ɜː/ → [øə]. However, the detrimental effect an MP word has on intelligibility has been questioned in the past with the argument that the members of MPs can usually be distinguished through context (Brown, 1989; Levis & Cortes, 2008). The facilitating effect of context on intelligibility extends beyond MP words and has been documented more generally in various studies (e.g., Garcia & Cannito, 1996; Hustad & Beukelman, 2001). Research question 3 therefore asks:

(RQ3) Is the (potential) difference in international intelligibility between words involving the substitution /ɜː/ → [øə] vs. non-MP words and MP words involving /æ/ → [ɛ] (i.e., /ɛ/) influenced by the availability of contextual support?

This question is of interest since the role played by contextual cues for international intelligibility is unclear. Jenkins (2000) and Deterding (2013) found pronunciation to be the major cause of communication breakdowns in ELF interactions, and Jenkins, with regard to her own data, attributed this to many NNSs' "[d]ifficulties [...] with top-down skills, particularly in relation to making use of contextual cues" and their resulting "over-reliance on bottom-up skills" (p. 20), viz. the acoustic signal. This seems to suggest that contextual cues are unlikely to have a facilitating effect on international intelligibility, but a study by Osimk (2009) provided evidence to the contrary, showing that NNSs benefit from sentence context when listening to English stimuli produced by other NNSs. Moreover, a small pilot study by Thir (2021) on the basis of elicited, interactive speech data illustrated how NNS ELF users draw on different types of linguistic and extra-linguistic context when processing another ELF user's accent. Thus contextual support is a variable meriting further attention in studies on international intelligibility, because it may constitute a confounding factor responsible for some of the inconsistencies in research findings discussed above.

One issue deserving of consideration when examining the intelligibility of MP words in context is the type of word in which the sound substitution results. Thus, the replacement /æ/ → /ɛ/ may result in a word with the same part of speech (POS) as the intended word (e.g., *pan* → *pen*), but it may also result in a word with a different POS (e.g., *bad* → *bed*). MPs consisting of words with the same POS are more common than one might assume, as a corpus study by Levis & Cortes (2008) revealed. Since MP words with the same POS are more likely to be confused even when syntactic cues are present (Levis & Cortes, 2008), the variable 'POS status' has been taken into account here with regard to the MP words. Interestingly, this variable is sometimes included in FL calculations (see Brown, 1988).

2. Method

This study used a quantitative, experimental approach to counteract the drawbacks of the *a posteriori approach* used in many previous studies on international intelligibility. The experiment compared the intelligibility of four different English word types spoken with an Austrian accent to an international audience:

a. /ɜː/ words (e.g., *bird*)
b. /æ/ non-MP[1] words (e.g., *flat*)
c. /æ/ MP 'different POS' words (e.g., *bad*, its MP counterpart being *bed*)
d. /æ/ MP 'same POS' words (e.g., *pan*, its MP counterpart being *pen*)

The words involved the sound substitutions discussed above, that is, /ɜː/ words were realized with [øə], whereas the vowel in /æ/ words was realized close to cardinal [ɛ]. Moreover, the post-vocalic /ɹ/ in the /ɜː/ words was not realized, going against the recommendations of the LFC (Jenkins 2000), which might be regarded as an additional threat to international intelligibility.

Word intelligibility was operationalized as correct identification of a word, using an exact word match orthographic transcription. To examine the impact of contextual factors on word intelligibility, the four word types were distributed evenly across four different listening conditions varying in the availability of contextual support (see Section 2.2.2.). The experiment used a within-subject design to reduce the impact of confounding factors arising from between-subject variables, i.e., each listener experienced all four conditions in a randomized order.

[1] The labels MP and non-MP here refer to the distinction /æ/-/ɛ/.

2.1 Participants

2.1.1 *Speaker*

A male Austrian NNS of English, aged 68 years, was recorded reading aloud the English words and sentences used as stimuli in this experiment (see Section 2.2.). He was selected because his accent exhibited the two sound substitutions of interest. His formal instruction in English lasted for 3 years, but, having used English professionally in business contexts for 40 years, he had ample experience using English in lingua-franca contexts. The recording was separated into individual files containing one test item each (i.e., a test word or a carrier sentence including a test word).

2.1.2 *Listeners*

508 NS and NNS listeners of English (male = 175, female = 330, other = 3) were recruited via e-mail, social media, and with the help of the author's international contacts. Their ages ranged from 18–74 years, but most (80%)[2] were between 18–35 years, resulting in a mean age of 29.4 years (median: 26 years). Fifty-eight participants (11%) identified as NSs, 434 (85%) as NNSs, and the remaining 16 participants (3%) chose the option 'I'm not sure'. They came from several different L1 backgrounds; some had two or three first languages. The most common L1 backgrounds are listed in Table 6 in the Appendix.

Participants' self-assessed familiarity with an Austrian accent in English was relatively low. Thirty-one percent reported no contact, and a further 37% said they had had 'very little' or 'rather little' so far, whereas 17% had had 'some' and the remaining 16% had had 'rather much' or 'very much'. Similarly, 35% indicated that they never hear the Austrian accent and a further 36% hear it only once a year or less. Nineteen percent had more regular exposure, hearing it several times a year or a month. The remaining 10% are exposed to it several times a week or daily.

Participants were also asked to estimate their listening proficiency in English using descriptions[3] that corresponded to the six proficiency levels of the *Common European Framework of Reference* (CEFR; see Table 1). The majority of participants (67%) believed they were highly proficient in English listening, estimating themselves to be either at C1 or C2 level. More than a quarter of the participants (29%) said they were at the intermediate or upper intermediate levels (i.e., B1 or B2), and only 4% regarded themselves as low-proficiency listeners (A1 or A2).

[2] All percentages in this section are rounded values, and may therefore not always add up to 100%.

[3] These were adaptations of the can-do statements constituting the CEFR's self-assessment scale for listening in a foreign language (2018, p. 167).

Table 1. Participants' self-assessed listening proficiency

CEFR level	n	%
A1	8	2
A2	12	2
B1	52	10
B2	95	19
C1	129	25
C2	212	42

2.2 Speech materials

2.2.1 *Target words*

Using the *Longman Pronunciation Dictionary* (Wells, 2008) as a reference guide, monosyllabic content words were selected containing one of the two target sounds (i.e., either /æ/ or /ɜː/) in both General American and Standard British pronunciation. To increase the likelihood that both NS and NNS listeners were familiar with the target words, only words labelled A1-B2 in the *English Vocabulary Profile* (EVP; Cambridge University Press, 2012) were included. The EVP provides information on the CEFR level at which L2 learners of English are able to use a word in a text. Since there is evidence that L2 learners' vocabulary knowledge is larger in terms of reception than production (e.g., Laufer, 1998), the range A1-B2 was considered adequate for the selection of target words even for participants of proficiency levels lower than B2. However, it is possible that some of the few lower-proficiency participants in the study were unfamiliar with certain target words, which may have affected the results.

For each of the four word types examined, there was one monosyllabic target word per condition, that is, there were 16 target words in the experiment (see Table 2). Each condition also included nine mono- or disyllabic distractor words containing different English vowels, and two disyllabic target words (one /ɜː/ and one /æ/ word) which were not part of the analysis presented here.[4]

A particular challenge was the choice of MP words. The pool of /æ/-/ɛ/ MPs in English is restricted to begin with, and moreover, the selection criteria described above were applied to *both* words in an MP, to increase the likelihood that it would constitute an MP from the point of view of NNS listeners. Thus, a

[4] The additional factor of word length exceeds the scope of this paper. The experiment's focus on monosyllabic words was intended to prevent a ceiling effect, since longer words can be more easily recognized in speech than shorter words (see Pisoni & McLennan, 2016).

very limited number of suitable MPs remained, some of which did not perfectly qualify as either 'different POS' or 'same POS', since one or both words in the MP exist as multiple POS in English. Their classification was based on considerations of whether L2 listeners might be more likely to perceive them as 'different POS' or 'same POS'. For example, *sand* in English is both a noun and a verb, but only the noun is included in the EVP (B1 level); there is no entry for the verb *sand*, which is rather infrequent apart from certain specialized contexts. Since the verb *sand* can be assumed to be unfamiliar to the average English L2 listener, the pair *sand-send* was classified as MP 'different POS'. However, both noun and verb meanings of *land* are included in the EVP at levels below C1, so *land-lend* may indeed be perceived as a 'same POS' MP by L2 listeners. The same is true for *gas–guess* (both the noun and the verb *guess* appear at levels below B2 in the EVP). In sum, the distinction 'same POS' and 'different POS' is not always clear-cut, and the classifications proposed here reflect tendencies rather than a perfect correspondence with either category.

Table 2. Target words in the experiment. For MP words, the /ɛ/ word is provided as well

/ɜː/	/æ/ non MP	/æ/ MP 'different POS'	/æ/ MP 'same POS'
birth	*rat*	*sand* (n) – *send* (v)	*land (v.) – lend (v.)*
nurse	*flat*	*bad* (adj.) – *bed* (n)	*gas* (n) – *guess* (n)
bird	*van*	*dad* (n) – *dead* (adj./adv.)	*pan* (n) – *pen* (n)
firm	*cab*	*bag* (n) – *beg* (v)	*pants* (plural n) – *pence* (plural n)

Note. N = noun, v = verb, adj. = adjective, adv. = adverb.

2.2.2 *Construction of test sentences*

To test the effect of contextual support on the intelligibility of the target words, four experimental conditions were devised. So far, 'context' has been used in its broad sense here, but when developing the experiment, a distinction was made between linguistic *co-text*, i.e., "the intratextual relations [between] linguistic elements" (Widdowson, 2019, p. 10), and extra linguistic *context*. The latter term is henceforth used in this restricted sense.

One of the four conditions tested the effect of syntactic co-text on word intelligibility (SYN condition), that is, to what extent purely grammatical information aided the identification of the target word. Words were embedded in short sentences indicating the POS of the target word:

(Ex a) *It's quite* ________. (flat)

Another condition was designed to test the effect of grammatical and semantic co-text on word intelligibility (SYN+SEM condition). Carrier sentences indicated the POS of the target word and contained a semantic cue in the form of a meaning relationship between the target word and a prime word. The prime (underlined in the example) always occurred before the target word.

(Ex b) *They found the <u>feather</u> of a* ______. (bird)

Yet another condition aimed at testing the effect of grammatical co-text and schematic information on word intelligibility (SYN+SCH condition). Schemata are knowledge structures based on previous experiences (Gureckis & Goldstone, 2011), which are regarded here as cognitive, extra-linguistic contexts that language users bring with them (cf. Widdowson, 2004). The condition was intended to activate a relevant schema in listeners by means of a 'schematic cue', operationalized as a short description of the situation under which the statement was made (in bold in the example below). This description was presented in writing to listeners prior to hearing the carrier sentence. Thus, extra-linguistic cognitive context was activated by means of additional co-text. Carrier sentences indicated the POS of the target word and connected meaningfully to the schematic cue, but did not contain any semantic primes as in the SYN+SEM condition.

(Ex c) ***When getting up in the morning***
I can't find my ________. (pants)

Finally, there was a control (C) condition, in which target words were presented without any co- or context. All words used in the carrier sentences in the other three conditions were A1-B2 levels according to the EVP and contained neither of the target sounds to avoid phonological priming effects. A full list of the carrier sentences is provided in Table 3.

2.3 Procedure

Upon accessing the experiment on the internet, participants were asked to adjust the volume of their computers[5] using a short test audio spoken in a Standard British accent. The audio contained neither of the target sounds. Participants were encouraged to use headphones, and asked to confirm that they were completing the experiment in calm surroundings, did not suffer from a hearing impediment, and were participating for the first time. After providing consent, they completed the four listening conditions in randomized order. They were unable to control

[5] Since the experiment was unsuitable for completion on mobile devices, individuals who tried to participate via such devices were denied access by means of a filter in the online system.

Table 3. Full list of target words and carrier sentences. Target words are underlined. Schematic cues are in small caps

	/ɜː/	**/æ/ non MP**	**/æ/ MP 'different POS'**	**/æ/ MP 'same POS'**
C condition	birth	rat	sand	land
SYN condition	There's a nurse.	It's quite flat.	It's quite bad.	It's a gas.
SYN+SEM condition	They found the feather of a bird.	He was driving a van.	He missed his relatives, especially his dad.	He fried the vegetables in a pan.
SYN+SCH condition	WHEN BUYING A NEW BED	ON THEIR WAY HOME FROM THE PUB	AT THE AIRPORT	WHEN GETTING UP IN THE MORNING
	This one is quite firm.	Let's get a cab.	I need to pick up my bag.	I can't* find my pants!

* Pronounced with /ɑː/.

the audio, which started playing automatically once a particular test item was accessed, preventing them from listening more than once.

In each of the four conditions, participants typed the missing word in the blank space provided. In the three conditions involving carrier sentences, participants saw the sentences on their screen with the target word replaced by a blank space. In the SYN+SCH condition, the schematic cue was displayed for five seconds on its own before the audio played and the carrier sentence appeared.

All four conditions were timed to avoid guessing and to increase task difficulty. From the onset of the audio, participants received a maximum of 11 seconds per item to enter their answer in the C and the SYN conditions and a maximum of 12 seconds per item in the SYN+SEM and the SYN+SCH conditions[6] (they could see the timer running and could use the 'next' button to proceed faster to the next item). Before the start of the trials, they received a timed practice item to develop familiarity with the task and the time constraint. They could control the start of the first trial by clicking a button on their screens. After completing all four conditions, participants filled in a questionnaire.

[6] Participants received one more second in the SYN+SEM and the SYN+SCH condition because here the audios were slightly longer than in the other two conditions. These time limits were carefully piloted with NNSs of English using different typing speeds.

3. Results

3.1 Effect of *word type* on intelligibility

A Friedman test comparing intelligibility across the four word types proved significant ($\chi^2(3)=589.727$, $p<.001$). Therefore, pairwise comparisons were performed with a Bonferroni correction for multiple comparisons. Effect sizes for each contrast on the basis of a Wilcoxon signed-rank test are given in Table 4.

Table 4. Effect sizes for each contrast between the four word types

Contrast		*r*
/ɜː/	/æ/ non-MP	.46
	/æ/ MP 'same POS'	.52
	/æ/ MP 'different POS'	.30
/æ/ non-MP	/æ/ MP 'same POS'	.23
/æ/ non-MP	/æ/ MP 'different POS'	−.26
/æ/ MP 'different POS'	/æ/ MP 'same POS'	.45

Note. Following Plonsky & Oswald (2014), $r=.25$ is considered a small effect, $r=.40$ a medium effect and $r=.60$ a large effect. These benchmarks are stricter than the ones originally proposed by Cohen (1988).

This revealed that /ɜː/ words were significantly more intelligible than each type of /æ/ words ($p<.001$ for each comparison; see Figure 1). However, the extent of this difference in intelligibility varied according to the type of /æ/ word, with the difference between /ɜː/ words and /æ/ non-MP words being medium ($r=.46$), whereas the one regarding /æ/ MP 'same POS' was medium to large ($r=.52$) and the one regarding MP 'different POS' words was small ($r=.30$). This was due to differences in intelligibility between the three types of /æ/ words examined. Interestingly, the /æ/ non-MP words were not the most intelligible among the three categories: while they were significantly more intelligible than the /æ/ MP 'same POS' words ($p<.001$, $r=.23$), they were significantly less intelligible than the /æ/ MP 'different POS' words ($p<.001$, $r=-.26$), though ultimately, these differences were small. Notably, the difference in intelligibility between the two types of MP words examined was not only significant but sizable ($r=.45$, $p<.001$).

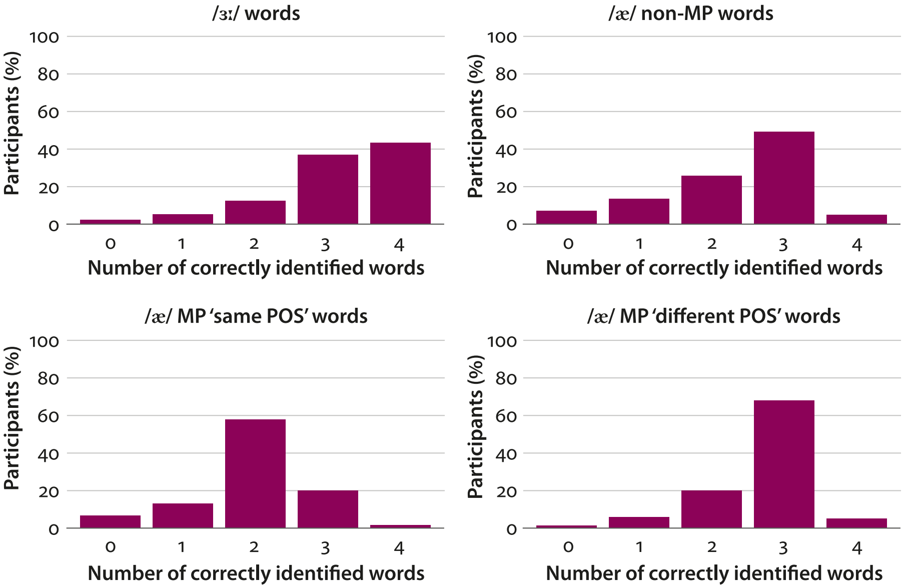


Figure 1. Barplots showing the proportions of participants who identified a certain number of target words correctly for each word type. The word type that was most intelligible was the /ɜː/ words (Mdn = 3, mode = 4), followed by the /æ/ MP 'different POS' words (Mdn = 3, mode = 3), the /æ/ non-MP words (Mdn = 3, mode = 3) and the /æ/ MP 'same POS' words (Mdn = 2, mode = 2)

3.2 Effect of *word type* X *condition* on intelligibility

To examine whether the identified differences in intelligibility between the four word types were equal across the four experimental conditions, a logistic mixed effects model was computed using the *glmer* function from the *lme4* package (Bates, Maechler, Bolker, & Walker, 2015) in R. Using a mixed effects model allowed me to account for the within-subject nature of the experiment, that is, the fact that observations were nested within participants. The optimizer *nlminbwrap* was used to avoid convergence errors. Correct identification of a word (yes/no) was entered as the Bernoulli distributed dependent variable. A model evaluation procedure was carried out, using a log-likelihood ratio test and the *Akaike's information criterion* (AIC; Akaike, 1974) to assess goodness-of-fit.

To construct the random effects structure of the model, a base model with a random intercept for subject (i.e., participant) was constructed. Since different items were used for a particular word type in all of the four conditions, a random intercept for item was added (cf. Baayen, 2013), which significantly improved the model fit, as did a by-subject random slope for conditions, which accounts for the fact that participants might react differently to the four conditions. It was also tested whether a by-subject random slope for sound would improve the model, thus accounting for the fact that certain subjects might react differently to the two different sound substitutions tested (e.g., due to varying levels of familiarity with an Austrian accent, or the influence of their L1 phonology). However, since this additional random slope specification did not significantly improve the model ($p = .106$) and only very marginally lowered the AIC by 0.1, it was not included in the final model.

Both condition and word type were entered as fixed effects into the model, as well as an interaction term for these two factors. While the factor condition and the interaction term significantly improved the model's fit ($p < .001$ for both), the factor word type itself improved it only slightly, just below the significance level ($p = .0598$). It was still retained because it improved the AIC from 6376.5 to 6375.1 and because the higher order interaction was significant. Once these predictors were entered into the model, the benefit of including a by-item random intercept disappeared, with the variance explained by this random effect now amounting to 0.00. It was therefore omitted from the random effects structure in the final model, which led to an improved AIC from 6293.9 to 6291.9. The by-subject random slope for conditions, however, still improved the model significantly at this point and was therefore retained.

The /ɜː/ word in the C condition was set as the reference category (intercept), that is, the final model evaluated to what extent the intelligibility of all other words differed to that of this word in this condition. Table 5 summarizes the results of the final model. There was a significant main effect of word type in the C condition, with each of the three /æ/ words being significantly less often understood ($p < .001$) than the /ɜː/ word. Condition also affected the intelligibility of the intercept significantly, with the /ɜː/ word being significantly more intelligible in the SYN, the SYN+SEM and the SYN+SCH condition than in the C condition ($p < .001$ for each comparison). Moreover, word type interacted significantly with condition in all possible combinations (mostly at $p < .001$, except for two combinations which were significant at $p < .05$ and $p < .005$, respectively). That is, the effect of word type on intelligibility vis-à-vis the intercept observed in the C condition was obviously not equal in the remaining three conditions.

By adjusting the reference category for condition and word type, p-values were obtained for the difference in intelligibility between the /ɜː/ word and the other three words in the remaining three conditions, as well as for the difference in intelligibility between the three /æ/ words in each condition. These were adjusted using Holm's (1979) method to correct for multiple comparisons. In addition, the *probability of being correctly identified* (PCI), as predicted by the final model, was computed for each word in each condition (see Figure 2). Based on

Table 5. Summary of the final model. The intelligibility of the /ɜː/ word in the C condition served as the reference category (intercept), to which the intelligibility of all other words is compared

Effect	Estimate	SE	z-value	p-value
Fixed effects				
Intercept	.034	.107	.322	.747
/æ/ MP 'different POS'	−3.064	.228	−13.439	<.001
/æ/ MP 'same POS'	−1.755	.168	−10.435	<.001
/æ/ non-MP	−2.904	.218	−13.309	<.001
SYN	2.813	.225	12.528	<.001
SYN+SEM	3.603	.273	13.214	<.001
SYN+SCH	2.108	.203	10.398	<.001
SYN x /æ/ MP 'different POS'	2.442	.313	7.790	<.001
SYN+SEM x /æ/ MP 'different POS'	2.288	.335	6.830	<.001
SYN+SCH x /æ/ MP 'different POS'	4.480	.328	13.663	<.001
SYN x /æ/ MP 'same POS'	−3.578	.345	−10.372	<.001
SYN+SEM x /æ/ MP 'same POS'	.643	.294	2.189	.029
SYN+SCH x /æ/ MP 'same POS'	1.902	.255	7.448	<.001
SYN x /æ/ non-MP	1.795	.302	5.938	<.001
SYN+SEM x /æ/ non-MP	.981	.324	3.031	.002
SYN+SCH x /æ/ non-MP	2.102	.284	7.396	<.001

Random Effects	Variance	Std.Dev.	Corr		
Subject (Intercept)	.907	.952			
SYN	.925	.962	−.10		
SYN+SEM	2.821	1.680	−.10	.89	
SYN+SCH	2.214	1.488	−.07	.70	.95

the estimates in Table 5, the PCI indicates the direction and size of the effect of word type on intelligibility in each condition.

For the C condition, that is, without any co-textual or contextual support, the model predicts a probability of 50.9% for the /ɜː/ word to be correctly identified, whereas it is only about 5% for the /æ/ non-MP and /æ/ MP 'different POS' word, and 15.2% for the /æ/ MP 'same POS' word. This difference in intelligibility between the /ɜː/ word and the three types of /æ/ word, which is statistically significant (*p* < .001 for each contrast), is considerable. Interestingly, in this condition

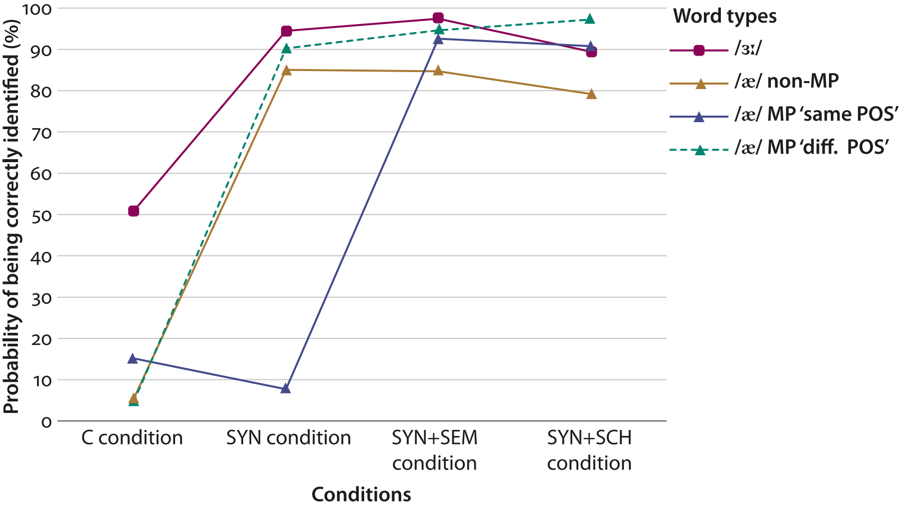


Figure 2. *Probability of being correctly identified* (PCI) for each word in each condition, as predicted by the final model

it does not seem to make much difference for intelligibility whether or not the /æ/ word is an MP word: the difference between the non-MP and the MP 'different POS' word is not significant ($p = .886$), and while the MP 'same POS' word is significantly more intelligible than the other two /æ/ words ($p < .001$ for both contrasts), the size of this difference is not dramatic.

In the SYN condition, the model predicts a 94.5% probability for the /ɜː/ word to be correctly identified. The PCI for the /æ/ non-MP and the /æ/ MP 'different POS' word is similarly high, amounting to 85.0% and 90.3%, respectively. The difference in intelligibility between the /ɜː/ word and these two /æ/ words is still statistically significant ($p < .001$ and $p = .018$), but it is far smaller than in the C condition. However, the difference in PCI between the /æ/ MP 'same POS' word and all other words is not only significant ($p < .001$ for each contrast), but quite extreme, since its PCI amounts to only 7.7%.

In the SYN+SEM condition, the PCI of all words is fairly high, in particular for the /ɜː/ word and the two MP words (97.4%, 94.6%, and 92.6%, respectively). Although the difference in intelligibility between the /ɜː/ word and the two MP words is significant ($p = .009$, $p < .001$), it is minor. Moreover, the difference between the two MP words is no longer significant ($p = .342$). The difference in intelligibility between the /æ/ non-MP word (PCI = 84.7%) and the other word types, which is significant at $p < .001$ for all three contrasts, is a little higher yet also not dramatic.

A similar picture emerges when considering the SYN+SCH condition: again, the PCI of all words is fairly high, but the /ɜː/ word is not the most intelligible in this condition, in contrast to all the others. The word with the highest PCI in this condition is in fact the /æ/ MP 'different POS' word (with 97.2%), followed by the/æ/ MP 'same POS' word (90.8%), the /ɜː/ word (89.5%), and the /æ/ non-MP word (79.2%). The difference in intelligibility between the /ɜː/ word and the /æ/ MP 'same POS' word is insignificant ($p=.886$), and the one between the /ɜː/ word and the two remaining /æ/ words is significant ($p<.001$ for both contrasts) yet not extreme. The latter also applies to the contrast between the /æ/ MP 'same POS' and the other two /æ/ words ($p<.001$ for both). The PCI of the /æ/ MP 'different POS' word exceeds that of the /æ/ non-MP word by 18.0%, a considerable difference in intelligibility which was statistically significant ($p<.001$).

4. Discussion

This study aimed to complement existing qualitative research findings regarding the role of /ɜː/ for international intelligibility on the basis of quantitative, experimental evidence. The first two research questions asked whether, in line with the LFC but contrary to FL predictions, words involving the substitution /ɜː/ → [øə] would be less intelligible to an international audience than both MP and non-MP words involving the substitution /æ/ → [e]. The results presented in Section 3.1. and 3.2. provide evidence to the contrary. The /ɜː/ words were significantly more intelligible than both /æ/ MP and non-MP words overall and in three of the four experimental conditions. This lends support to the explanatory potential of FL regarding international intelligibility, and suggests that the special status accorded by the LFC to /ɜː/ vis-à-vis other English vowels is not tenable when the substitution of /ɜː/ is phonetic rather than phonemic. This finding is also in line with Munro & Derwing's (1995) study using NS listeners, which showed that phonetic errors did not correlate with reduced intelligibility while phonemic errors did to some extent.

RQ3 addressed the impact of contextual support on intelligibility between the different word types examined. Here, the experiment revealed several interesting observations. In the absence of any co-textual or contextual support, the replacement /ɜː/ → [øə] turned out to be far more intelligible than the replacement /æ/ → [ɛ], regardless of whether /æ/ non-MP words or the two types of MP words examined ('same POS' and 'different POS') were concerned. However, once syntactic co-text was available, the difference in intelligibility between the replacement /ɜː/ → [øə] and the replacement /æ/ → [ɛ] decreased dramatically. This is even more notable in the light of the fact that the carrier sentences in the SYN condition were very short and simple, yet they almost eliminated the difference in intelligibility between the two vocalic substitutions and led to almost perfect intelligibility of the target words. The notable exception to this 'equalizing' and 'neutralizing' effect of syntactic co-text was the /æ/ MP 'same POS' word, which was hardly intelligible in the SYN condition. Since the replacement /æ/ → [ɛ] here resulted in a word of the same lexical category, listeners could not disambiguate the intended word and the word they heard, despite the available co-textual support.

Since the different types of target words (apart from the /æ/ MP 'same POS' word) were already highly intelligible in the SYN condition, it is difficult to assess the impact of the additional semantic and schematic cues provided in the remaining conditions: clearly, for most target words, there was not much to be gained in terms of intelligibility once they were presented within simple syntactic co-text. This ceiling effect is surprising considering that participants worked under a time limit, but it might be because most were highly proficient listeners in English. However, some interesting observations can be made. Similar to the syntactic co-text in the SYN condition, the additional semantic cue in the SYN+SEM condition worked as both an equalizer and neutralizer by dramatically increasing the intelligibility of the /æ/ MP 'same POS' word, bringing it to an intelligibility level comparable to that of the other words. It also further increased the intelligibility of the /ɜː/ word and the /æ/ MP 'different POS' word and decreased their difference in intelligibility. However, it did not seem to contribute positively to the intelligibility of the /æ/ non-MP word. One possible explanation is that the semantic cues in this condition may have exhibited different levels of predictiveness regarding the target word, that is, for some reason, listeners experienced the one in the carrier sentence containing the /æ/ non-MP word to be substantially less predictive than the others. Although I attempted to create cues of comparable predictiveness, and the items were piloted in advance, certain differences in predictiveness will always remain. The same is true for the schematic cues provided in the SYN+SCH condition, which bear the additional limitation that schemata are "social constructs" (Widdowson, 2004, p. 43) that are even more elusive than a semantic relationship encoded in the language. Creating schematic cues of comparable predictiveness for a sample as culturally heterogeneous as the one examined here can thus only be attempted but probably never be fully achieved. This might explain why in two cases (one with the /ɜː/ word), the additional schematic cue did not contribute positively to the intelligibility of the target word. However, similar to the semantic cue in the SYN+SEM condition, the schematic cue substantially increased the intelligibility of the /æ/ MP 'same POS' word. Both the semantic and the schematic cues were crucial in disambiguating the MP word from its same-POS counterpart, which syntactic co-text alone as provided in the SYN condition was unable to do.

Summing up, a FL effect was clearly observable when words were heard in isolation, but was noticeably attenuated (or eliminated) when co-textual and contextual supports were available. Thus, this study suggests that co-textual and contextual information may profoundly influence word intelligibility to international listeners, even overriding the effects of FL, at least if listeners are highly proficient. This finding is inconsistent with that of Jenkins (2000), who attributed an overreliance on bottom-up processing strategies – and thus a limited ability to benefit from co-textual and contextual cues – even to NNS listeners “at upper-intermediate level and beyond” (p. 83). However, real world conditions of processing language differ from those in the current experiment, in which participants processed one word or sentence at a time and where co- and contextual support was provided in written form. In actual ELF interactions, participants might not be able to draw on co-text and context to the same extent, due to processing overload or because parts of the surrounding co-text of a word are unintelligible. Still, the results of this study indicate that co-text and context are important variables that should receive greater attention in research on intelligibility in ELF communication.

5. Limitations and suggestions for further research

There are a few limitations to this study that should be addressed in further research. The first relates to the influence of sound quality and background noise. Though the use of headphones was encouraged and participants were asked to use ‘calm surroundings’, the researcher had no control over the conditions under which listeners were taking part. This trade-off was accepted considering that the online experiment resulted in a large, varied international sample including listeners from many parts of the world who were entirely unfamiliar with the Austrian accent in English. However, further studies should control for these potential confounding factors.

Another limitation was the use of self-reports to assess participants’ listening proficiency, an approximate measure, since language users may either under- or over-estimate their language skills (e.g., Trofimovich et al., 2016). Although the CEFR ‘can-do’ statements constitute a more objective self-assessment tool than proficiency-scales ranging from ‘excellent’ to ‘poor’, they do not provide the same degree of objectivity as standardized test scores. If such measurements can be obtained from participants, they are preferable.

A further limitation concerns the sample of this study. Although some participants may have overestimated their listening proficiency in English, most seemed to be highly proficient listeners. However, lower-proficiency listeners may be unable to rely on co- and contextual cues to the same extent. The results are therefore only generalizable to advanced English listeners, and should be complemented by further research addressing the differences in processing strategies between NNS listeners at different proficiency levels.

6. Conclusion

This study provides evidence for the explanatory potential of FL to issues of international intelligibility. It also suggests that the LFC's recommendation to prioritize the teaching of /ɜː/ over other English vowel qualities is inappropriate when learners employ a phonetic rather than a phonemic substitution. Clearly, the LFC is best viewed as an "ongoing empirical description" (Walker, 2010, p. 44) rather than a monolithic set of prescriptions for how to maintain international intelligibility, and its general pedagogic recommendations should be adapted in the light of new research findings on how different types of sound substitutions affect international intelligibility.

Moreover, this study showed that FL is only one part of the bigger puzzle that is international intelligibility. The observed 'neutralizing' effect of co-textual and contextual cues regarding the detrimental impact of certain pronunciation features, viz. the impact of FL, indicates the powerful influence that such factors could have in international contexts. This points to the need for a reconsideration of the focus of ELF intelligibility research: rather than seeking to identify the pronunciation features that contribute most to intelligibility problems in international contexts, it might be more important to determine the co-textual and contextual conditions under which certain features – or pronunciation in general – become crucial for international intelligibility. Such research has important implications for teaching: whereas learners may sometimes converse under conditions where there is ample co- and contextual information to compensate for pronunciation errors (e.g., in face-to-face conversations on familiar topics), they may at other times use ELF in situations where such information is scarce, and accurate pronunciation is therefore of much greater importance for mutual intelligibility (e.g., when dictating an address or a list via the telephone). Teachers should raise learners' awareness for such differences in co-textual and contextual conditions, that is, when they will have to pay particular attention to their pronunciation to remain internationally intelligible.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Baayen, R. H. (2013). Multivariate statistics. In R. J. Podesva & D. Sharma (Eds.), *Research methods in linguistics* (pp. 337–372). Cambridge, UK: Cambridge University Press.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Berger, A. (2010). German. In R. Walker, *Teaching the pronunciation of English as a lingua franca* (pp. 107–110). Oxford, UK: Oxford University Press.

Brown, A. (1988). Functional load and the teaching of pronunciation. *TESOL Quarterly*, 22(4), 593–606. https://doi.org/10.2307/3587258

Brown, A. (1989). Some thoughts on intelligibility. *The English Teacher*, XVIII. http://www.melta.org.my/ET/1989/main4.html

Cambridge University Press. (2012). *English vocabulary profile*. http://vocabulary.englishprofile.org/staticfiles/about.html

Catford, J. C. 1987. Phonetics and the teaching of pronunciation. In J. Morley (ed.) *Current perspectives on pronunciation* Washington, D.C.: (pp. 87–100) Teachers of English to Speakers of Other Languages.

Chan, A. Y. W. & Li, D. C. S. (2000). English and Cantonese phonology in contrast: Explaining Cantonese ESL learners' English pronunciation problems. *Language, Culture and Curriculum*, 13(1), 67–85. https://doi.org/10.1080/07908310008666590

Cohen, J. W. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cole, S. (2002). An investigation of the role of vowel quality in oral interactions between NNSs of English as an international language. *Speak Out!*, 29, 28–37.

Collins, B., & Vandenbergen, A.-M. (2000). *Modern English pronunciation: A practical guide for speakers of Dutch*. Gent, BE: Academic Press.

Council of Europe. (2018). *Common European framework of reference for languages: Learning, teaching, assessment* (Companion volume with new descriptors). https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

Derwing, T. M. & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, 19(1), 1–16. https://doi.org/10.1017/S0272263197001010

Derwing, T. M. & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. Amsterdam: John Benjamins. https://doi.org/10.1075/lllt.42

Deterding, D. (2013). *Misunderstandings in English as a Lingua Franca: An analysis of ELF interactions in South-East Asia*. Berlin: de Gruyter Mouton. https://doi.org/10.1515/9783110288599

Deterding, D., & Kirkpatrick, A. (2006). Emerging South-East Asian Englishes and intelligibility. *World Englishes*, 25(3–4), 391–409. https://doi.org/10.1111/j.1467-971X.2006.00478.x

Garcia, J. M. & Cannito, M. P. (1996). Influence of verbal and nonverbal contexts on the sentence intelligibility of a speaker with dysarthria. *Journal of Speech, Language, and Hearing Research*, 39(4), 750–760. https://doi.org/10.1044/jshr.3904.750

Gureckis, T. M., & Goldstone, R. L. (2011). Schema. In P. C. Hogan (Ed.), *The Cambridge encyclopedia of the language sciences* (pp. 725–727). Cambridge, UK: Cambridge University Press.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70.

Hustad, K. C., & Beukelman, D. R. (2001). Effects of linguistic cues and stimulus cohesion on intelligibility of severely dysarthric speech. *Journal of Speech, Language, and Hearing Research*, 44, 497–510. https://doi.org/10.1044/1092-4388(2001/039)

Jenkins, J. (2000). *The phonology of English as an international language*. Oxford, UK: Oxford University Press.

Kennedy, S. (2012). When non-native speakers misunderstand each other: Identifying important aspects of pronunciation. *Contact*, 38(2), 49–62.

Kim, H. & Billington, R. (2018). Pronunciation and comprehension in English as a lingua franca communication: Effect of L1 influence in international aviation communication. *Applied Linguistics*, 39(2), 135–158. https://doi.org/10.1093/applin/amv075

Komar, S. (2017). The relationship between the perception and production of four General British vowels by Slovene university students of English. *Linguistica*, 57(1), 161–170. https://doi.org/10.4312/linguistica.57.1.161-170

Laufer, B. (1998). The development of passive and active vocabulary in a second language: Same or different?. *Applied Linguistics*, 19(2), 255–271. https://doi.org/10.1093/applin/19.2.255

Levis, J. M. & Cortes, V. (2008). Minimal pairs in spoken corpora: Implications for pronunciation assessment and teaching. In C. A. Chapelle, Y. R. Chung & J. Xu (Eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 197–208). Ames, IA: Iowa State University.

Luchini, P. L., & Kennedy, S. (2013). Exploring sources of phonological unintelligibility in spontaneous speech. *International Journal of English and Literature*, 4(3), 79–88.

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, 45(1), 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

O'Connor, J. D. (1980). *Better English pronunciation* (2nd ed.). Cambridge, UK: Cambridge University Press.

O'Neal, G. (2015). ELF intelligibility: The vowel quality factor. *Journal of English as a Lingua Franca*, 4(2), 347–358. https://doi.org/10.1515/jelf-2015-0026

Osimk, R. (2009). Decoding sounds: An experimental approach to intelligibility in ELF. *Vienna English Working Papers*, 18(1), 64–89.

Pisoni, D. B., & McLennan, C. T. (2016). Spoken word recognition: Historical roots, current theoretical issues, and some new directions. In G. Hickok & S. L. Small (Eds.), *Neurobiology of language* (pp. 239–253). Tokyo: Academic Press. https://doi.org/10.1016/B978-0-12-407794-2.00020-1

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, 64(4), 878–912. https://doi.org/10.1111/lang.12079

Richter, K. (2019). *English-medium instruction and pronunciation: Exposure and skills development*. Bristol: Multilingual Matters. https://doi.org/10.21832/RICHTE2456

Sewell, A. (2017). Functional load revisited: Reinterpreting the findings of 'lingua franca' intelligibility studies. *Journal of Second Language Pronunciation*, 3(1), 57–79. https://doi.org/10.1075/jslp.3.1.03sew

Thir, V. (2021). The role of co-textual and contextual cues for intelligibility in ELF interactions. In A. Mauranen & S. Vetchinnikova (Eds.). *Language change: The impact of English as a Lingua Franca*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108675000.014

Trofimovich, P., Isaacs, T., Kennedy, S., Saito, K., & Crowther, D. (2016). Flawed self-assessment: Investigating self- and other-perception of second language speech. *Bilingualism: Language and Cognition*, 19(1), 122–140. https://doi.org/10.1017/S1366728914000832

Walker, R. (2010). *Teaching the pronunciation of English as a lingua franca*. Oxford, UK: Oxford University Press.

Wells, J.C. (2008). *Longman pronunciation dictionary* (3rd ed.). Harlow, UK: Pearson Education Longman.

Widdowson, H.G. (2004). *Text, context, pretext: Critical issues in discourse analysis*. Malden, MA: Blackwell. https://doi.org/10.1002/9780470758427

Widdowson, H.G. (2019). *On the subject of English: The linguistics of language use and learning*. Berlin: de Gruyter Mouton. https://doi.org/10.1515/9783110619669

Zoghbor, W. (2010). The effectiveness of the Lingua Franca Core (LFC) in improving the perceived intelligibility and perceived comprehensibility of Arab learners at post-secondary level (Doctoral thesis, University of Leicester, Leicester, UK). Retrieved from https://pdfs.semanticscholar.org/fbcf/6dcb0eb0ed58a5a02222253c603a0776b394.pdf

Appendix.

Table 6. L1 backgrounds in the sample

L1 background	Rounded % of total number of participants
Turkish	10
English	9
Italian	8
Portuguese	5
French	5
Greek	4
Arabic	4
Russian	4
Finnish	3
Thai	3
Danish	3
Persian	3
Serbo-Croatian	3
Chinese (Mandarin)	3

Table 6. *(continued)*

L1 background	Rounded % of total number of participants
Japanese	3
Hungarian	2
Spanish	2
Dutch	2
Polish	2
Catalan, Spanish	1
Chinese (Cantonese)	1
Chinese (other/unspecified)	1
German	1
Swedish	1
Remaining L1 backgrounds	16

Note. No exact figures are given for L1 backgrounds making up less than 1% of the sample (unrounded value). Different combinations of two or three first languages were counted as different L1 backgrounds. The category ‘Remaining L1 backgrounds’ also includes several combinations involving English.

Investigating the relationship between comprehensibility and social evaluation[*]

Charlotte Vaughn and Aubrey Whitty
University of Oregon

The processing fluency hypothesis proposes that listeners' perceived difficulty processing the speech of L2 speakers (called *comprehensibility/processing fluency*) leads them to downgrade those speakers socially. In this paper, we investigate this relationship, focusing on context-specificity. L1-English listeners provided comprehensibility and social evaluation ratings of L1-Korean speakers speaking English, while an orthographic depiction of the speech either appeared alongside the audio or did not, a manipulation aiming to affect comprehensibility. Varying orthography between subjects, Experiment 1 found that orthography resulted in greater comprehensibility, but not more positive social evaluations. Experiment 2 manipulated orthography within subjects, varying context: orthography trials were presented first or last. Comprehensibility and social evaluation ratings were related only when orthography was first, suggesting a conditional, asymmetrical relationship where listeners more readily downgrade than upgrade the same speaker when orthography changes. Our results highlight the context-dependent nature of these constructs, limiting the generalizability of the processing fluency hypothesis.

Keywords: social evaluation, status, solidarity, comprehensibility, processing fluency

1. Introduction

Decades of work on comprehensibility, accentedness, and intelligibility has demonstrated that these constructs are related but partially independent. For example, a speaker can be completely intelligible but rated as heavily accented (Derwing & Munro, 1997; Munro & Derwing, 1995a). From the perspective of

[*] An earlier version of this article was published as part of a special issue of the *Journal of Second Language Pronunciation* 6(3). https://doi.org/10.1075/jslp.20022.vau

https://doi.org/10.1075/bct.121.10vau

improving social outcomes for L2 learners, the insight that these constructs are not fully correlated has enabled emphasis on the role of the listener, moving a portion of the responsibility for successful communication off the shoulders of L2 speakers (Lippi-Green, 2012).

Accordingly, there has been a growth in scholarly interest in listener and contextual contributions to these constructs, acknowledging that they are not static properties of speakers, but rather emergent from combinations of speakers, listeners, and contexts (Derwing & Munro, 2009; Gluszek & Dovidio, 2010; Lindemann & Subtirelu, 2013). For example, several non-speaker factors influence one or more of these constructs, including listeners' social expectations about the speaker (Kang & Rubin, 2009; Vaughn, 2019), the congruence between expectations and the incoming signal (McGowan, 2015), listeners' attitudes toward social groups (Hu & Lindemann, 2009; Ingvalson et al., 2017a), listeners' cognitive characteristics (Ingvalson et al., 2017b), and the context in which the speech is rated (Nagle, Trofimovich, & Bergeron, 2019; Tzeng et al., 2016; Vaughn & Baese-Berk, 2019).

Another type of judgment, one that has substantial consequences for everyday interaction, is listeners' social evaluations of L2 speakers (see Giles & Rakić, 2014 for a review). Studies investigating social evaluation typically describe two empirically-supported dimensions, *status,* indexing competence, and *solidarity*, indexing warmth. Listeners' evaluations of non native speakers and speakers of marginalized native varieties tend to be negative, particularly along status dimensions (see Gluszek & Dovidio, 2010 for a review). These negative evaluations can have tangible consequences, for example, in the workplace (Cardoso et al., 2019; Hosoda & Stone-Romero, 2010) and in the courtroom (Dixon & Mahoney, 2004). As with comprehensibility, accentedness, and intelligibility, several contextual factors influence social evaluations, including the covert/overt prestige of the variety (Lindemann, 2003; Ryan, Giles, & Sebastian, 1982), the listener's relationship with the speaker (Bresnahan et al., 2002), task order (Brennan & Brennan, 1981), and the relative in-group status of the speaker with respect to a reference frame (Dragojevic & Giles, 2014).

It is of interest theoretically, pedagogically, and for the lives of L2 learners to better understand how social evaluation is related to comprehensibility, accentedness, and intelligibility. Some work across subfields has begun to bring these areas together. One particular focus of this work is the potential relationship between listeners' social evaluations of speakers and listeners' subjective experiences of task difficulty, which tends to be described as *comprehensibility* in language-related fields (Derwing & Munro, 1997; Munro & Derwing, 1995a) and as *processing fluency* in psychology (Alter & Oppenheimer, 2009). However, studies using the term 'comprehensibility' and those employing 'processing fluency' have largely proceeded in parallel, with little conversation between the literatures, even though these phenomena are nearly identical (see Munro, 2018). Each construct is operationalized by asking listeners to report how much effort was required to perform a task (e.g., listening to speech), or how difficult the task seemed. Hereafter, we discuss the construct of comprehensibility/processing fluency as *CPF* to acknowledge the synonymy of the two terms (and to avoid confusion with other uses of the term 'fluency' in the L2 learning literature).

1.1 Social evaluation and comprehensibility/processing fluency

Different lines of work investigating the relationship between CPF and social evaluation have assumed different directions of causality. Figure 1 depicts two opposing accounts. Route a posits that negative social evaluations *are a cause of* lower comprehensibility (or intelligibility in some studies; for a review see Lindemann & Subtirelu, 2013). For example, Taylor Reid, Trofimovich, and O'Brien (2019) found that priming Canadian English listeners with positive or negative comments about L1 French-L2 English speakers' English skills affected comprehensibility ratings. Ingvalson et al. (2017b) found that listeners' attitudes toward foreign-accented talkers significantly contributed to models predicting intelligibility scores. Further, the *reverse linguistic stereotyping* account (RLS, Kang & Rubin, 2019) endorses this direction of causality. RLS posits that a listener's expectation about a speaker's social category membership (e.g., an L2 speaker) and related stereotypes, potentially triggered by perceived accent strength (Ryan et al., 1977, dashed line in Route a), invokes social evaluations based on that category (e.g., more negative status ratings), and results in a listener inferring that the speaker is difficult to process and understand. A listener's degree of RLS is measured by how much their social evaluation of the speaker predicts comprehensibility or intelligibility (Kang & Rubin, 2019). In Route a, then, negative social evaluations of L2 learners are due to social categorization and stereotypes, and can result in decreased comprehensibility.

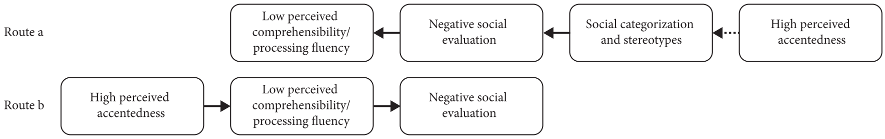

Route a
Low perceived comprehensibility/ processing fluency
Negative social evaluation
Social categorization and stereotypes
High perceived accentedness
Route b
High perceived accentedness
Low perceived comprehensibility/ processing fluency
Negative social evaluation

Figure 1. Two routes assumed in existing literature
Note. Routes a and b depict different causal directions assumed to underlie the relationship between comprehensibility/processing fluency and social evaluation of L2 speakers in prior studies.

Meanwhile, other work has suggested the opposite causal direction for this relationship, often discussed as the *processing fluency hypothesis* (or *fluency principle*), in which negative social evaluations of L2 speakers *are caused by* the listener's experience of processing disfluency (i.e., low comprehensibility), depicted in Figure 1, Route b. Building on work in social psychology (Alter & Oppenheimer, 2009; Lick & Johnson, 2015), this account suggests that greater perceived effort in processing a stimulus leads the perceiver to assign a more negative social evaluation to the stimulus because they interpret their increased effort as resulting from the stimulus itself. In Route b, then, negative social evaluations of L2 learners are due to misattributed processing difficulty.

The processing fluency hypothesis has been used to explain why L2 speakers are often downgraded in terms of status and/or solidarity traits (Dragojevic & Giles, 2016; Dragojevic et al., 2017; Dragojevic, 2020) and are judged as less credible than native speakers (Lev-Ari & Keysar, 2010). Proponents of this hypothesis acknowledge that listeners' stereotypes about social groups likely also contribute to negative social evaluations of L2 speakers, but offer processing fluency as separate from social/ideological factors (Dragojevic et al., 2017). These studies attempt to remove social/ideological factors from the equation (by holding them constant, for example), in an effort to isolate the role of cognitive factors like processing fluency.

Empirical support for this hypothesis has been mixed. For example, although Lev-Ari & Keysar (2010) found that listeners rated trivia statements to be more truthful when recited by a native than a nonnative speaker, which they attributed to the processing difficulty involved in processing nonnative speech, Souza & Markman (2013) failed to find such an effect using a range of similar designs (see also De Meo et al., 2011 and Stocker, 2017 for replication failures). In an in-depth examination, Ogden (2019) found no evidence for improved social evaluations with increases in an objective measure of comprehensibility, pupil dilation.

Several papers by Dragojevic and colleagues in support of the processing fluency account are direct points of departure for the present study. Dragojevic & Giles (2016) manipulated the presence or absence of background noise when presenting listeners with a speaker who performed a native English or Punjabi accent, and found that noise did not always affect status and solidarity ratings of either guise. Dragojevic et al. (2017) used accentedness to manipulate CPF (non-dashed line between accentedness and CPF in Route b) and found listeners who heard a performed "heavy" accent (Punjabi- or Mandarin-accented English) rated the speaker with less comprehensible/fluent CPF ratings and more negative status ratings, but with equivalent solidarity ratings as the "mild" performance of the same accent. Finally, Dragojevic (2020) presented listeners with either subtitles (orthography) or not alongside Mandarin-accented English speech from two speakers. Subtitles were correlated with increased status ratings in two of three experiments and increased solidarity ratings in one of three.

Ryan, Carranza, & Moffie (1977, see also Brennan & Brennan, 1981), in a study similar to Dragojevic et al. (2017), found that social evaluations of speakers became more negative as they were judged to be more accented. However, the authors interpreted these results in the opposite causal direction as Dragojevic et al.'s interpretation, suggesting instead that more accented speech triggered negative stereotypes, leading to more negative social evaluation (Route a). Thus, although the literature supports some relationship between social evaluations and judgments of speech, the direction of causation is not yet clear.

1.2 The present study

This paper focuses on the processing fluency hypothesis, with Experiment 1 serving as a replication and extension of prior studies, and Experiment 2 offering a context manipulation to test the generalizability of the hypothesis. We manipulate listeners' CPF by presenting orthographic text, or not, alongside the speech recording. As described above, most prior work in this area has taken fluently processed stimuli and degraded them in some way, such as embedding speech in noise (Dragojevic & Giles, 2016; Souza & Markman, 2013). Rather than decreasing listeners' CPF, we instead aim to *increase* listeners' ratings by adding orthography, an approach conducted in only one prior study (Dragojevic, 2020). If listeners can read what speakers are saying as they hear it, will they socially evaluate the speakers more positively, as predicted by the processing fluency account? Using orthography to increase CPF, rather than using noise to decrease it, takes into account the finding that strategies for processing speech in noise and accented speech are not identical (McLaughlin et al., 2018). Further, this approach ensures that low comprehension does not contribute to results, as it might if degrading the signal with noise.

We designed Experiment 1 as a near-direct replication of Dragojevic et al. (2017), with the addition of the crucial orthography manipulation.[1] We extend this prior work by using a different native language background (L1 Korean), and using two speakers rather than one to test orthography's interaction with levels of accentedness.

Experiment 2 is the first known within-subjects test of the CPF manipulation. This design controls the contexts of listeners' experiences of the speakers and the task, asking whether listeners' experiences over the course of the experiment (by

[1] Our model was this paper rather than Dragojevic (2020); the present study was conducted concurrently with this more recent study.

blocking trials by orthography condition) affect social evaluations. This allows us to ask whether any observed CPF effect holds across the board, or whether it is conditioned by orthography block order.

Although changes in orthography are not necessarily a common real world context, our task-based context manipulation is meant as an instance of the context-dependence that is always present in real world settings. Proponents of the processing fluency account have raised the important observation that social evaluations may have a cognitive basis in addition to having social/ideological origins. However, attempting to completely separate cognitive from ideological origins obscures the reality that cognitive factors are themselves situated within contexts, including social ones (Sumner et al., 2014), making it especially important to consider context. We return to this point, and the implications of our design for broader societal applications, in Section 4.

2. Experiment 1

2.1 Materials and methods

2.1.1 *Materials*

Stimuli were two short reading passages from the Wildcat Corpus (Van Engen et al., 2010), the Stella passage and the North Wind and the Sun (NWS) passage, recorded by two native Korean speakers. Although Dragojevic et al.'s (2017) study used one speaker who produced "heavy" and "light" accented guises of the same passage, here we used two speakers who we normed separately for accent ratings, since we did not want to assume correlation between accentedness and comprehensibility (Derwing & Munro, 1997). We conducted a small pilot study ($N=7$) to collect baseline accent ratings (1 = no accent at all, to 9 = very strong accent) for 8 male L1 Korean Wildcat Corpus speakers reading both passages. Two speakers differing in average accent ratings were selected, speaker K03 (age = 18) as the "more accented" (MA) speaker (M accent rating = 6.9), and speaker K17 (age = 30) as the "less accented" (LA) speaker ($M=2.9$). Average recording length was shorter for the LA speaker ($M=42$ sec) than the MA speaker ($M=53$ sec), and the Stella passage ($M=36$ sec) was shorter than the NWS passage ($M=59$ sec). Recordings were amplitude normalized to yield an approximately equal volume across stimuli.

2.1.2 *Procedure*

The experiment was conducted remotely using a web browser-based presentation software. Participants heard two passages (Stella and NWS), one produced by each speaker (MA and LA), and responded to the same series of questions for each passage. The order of presentation of speakers, passages, and speaker-passage combinations was counterbalanced. In the crucial manipulation (+/−orthography), participants were randomly assigned to the no accompanying orthography (*−orth*) or the accompanying orthography (*+orth*) condition.

Participants were told they would hear two passages and that for each they would be asked questions "about the speaker you listened to" and "about what they said." Participants were asked to wear headphones. After completing an audio check, the experiment began. Following Dragojevic et al. (2017), listeners were told they would be hearing "a male speaker from Korea" as an attempt to equate social categorization-driven stereotypes. In the *+orth* condition, participants were told they would see on the screen the text of the passage that the speaker was reading, and were asked to read along as they listened. Participants could listen to each passage only once, and could not advance until the passage had finished playing. After listening, seven blocks of questions (described below) were presented, always in the same order. Question order within blocks was randomized when applicable. Although our central concern was the relationship between CPF and social evaluation, we followed Dragojevic et al. (2017) closely and thus included all question blocks from that study in the design.[2] Below we describe and report only methods and results from CPF and social evaluation ratings; see Dragojevic et al. (2017) for further methodological details.

To measure social evaluation, listeners were asked to think about the speaker they just heard, and using a 7-point scale (1 = not at all and 7 = very), rate them on the following dimensions: successful, intelligent, smart, educated, competent (*status traits*), pleasant, nice, sociable, honest, and friendly (*solidarity traits*). Following Dragojevic et al. (2017), a status score was calculated by averaging the five status scales, and a solidarity score by averaging the five solidarity scales. To measure CPF, three questions were asked: How easy was the speaker to understand? (1 = not easy at all, to 7 = very easy), How clear was the speaker? (1= not clear at all, to 7 = very clear), How comprehensible was the speaker?[3] (1= not

[2] After an initial cloze test, blocks included questions about (in the following order): listener affect, speaker status, speaker solidarity, CPF, accentedness, where the speaker was from, prototypicality of speaker, societal attitudes toward Korean-accent, open-ended comments, and how many speakers they heard (the final two being our own additions).

[3] Following Dragojevic et al. (2017), we did not explicitly define comprehensibility, or any of the scales, for participants. Thus, we acknowledge that there is variability in how the scales were interpreted across participants, and that this particular use of the term comprehensibility could

comprehensible at all, to 7 = very comprehensible), again averaged to obtain an overall CPF score.[4] After the experiment, participants completed a demographic survey, including questions about their language background and exposure to Korean-accented English.

2.1.3 *Participants*

A total of 378 participants took part in Experiment 1. All were undergraduates recruited from the University of Oregon Human Subjects Pool. They participated online, and were given partial course credit for their time. Sixteen participants were excluded because they learned English after age 5, leaving 362 native or near-native English-speaking participants. Following Dragojevic & Giles (2016) and Dragojevic (2020), all trials where participants did not accurately identify that the speaker was from Korea were excluded, resulting in a removal of 27.3% of trials (equally distributed across +*orth* and −*orth* conditions), and a total of 47 participants.[5] This step left 315 participants (ages 18–39, *M* age = 20; 205 female, 103 male, 7 non-binary) with at least one trial for analysis (total trials = 526). The majority of listeners reported having infrequent exposure to Korean-accented English, 62.2% hearing the accent infrequently or very infrequently, 22.9% neither often nor infrequently, and 14.9% often or very often; this information did not improve any model fits and thus is not analyzed further.

2.2 Results

Statistical analysis was conducted with linear mixed effects models using the lmerTest package in R (Kuznetsova et al., 2017), with participant as a random intercept. In addition to the critical independent variables (speaker and +/−orthography), all experimental design and counterbalancing factors (passage, speaker order, passage order) were included as main effects. Two- and three-way interactions were included only if they significantly improved the model via likelihood ratio testing. All continuous measures (CPF score, status score, solidarity score) were centered and scaled for statistical analysis, but raw values are presented in figures and when reporting group means.

be understood differently from Munro & Derwing's intended meaning. We note that results obtained for the "ease of understanding" scale alone, perhaps most aligned with Munro & Derwing's use of comprehensibility, are comparable to results for the composite CPF score.

[4] In following Dragojevic et al. (2017) we used 7-point scales, although 9-point scales are suggested as best practice (Munro, 2018).

[5] For Experiment 1 and Experiment 2, major results were qualitatively similar regardless of whether this subset of the data or the complete dataset was used.

2.2.1 *Manipulation check*

The speaker and +/−orthography conditions were designed to influence CPF. Thus, we first determined whether CPF ratings differed across these conditions (see Figure 2a). Speaker significantly affected CPF ($\beta = 0.38$, $SE = 0.09$, $p < .001$), with the LA speaker ($M = 3.96$) rated with higher (more comprehensible/fluent) CPF than the MA speaker ($M = 3.24$). Orthography also significantly affected CPF ($\beta = 0.37$, $SE = 0.11$, $p < .0001$), with *+orth* given higher CPF ratings ($M = 3.98$) than *−orth* ($M = 3.23$). There was also a significant interaction between +/−orthography and speaker ($\beta = 0.31$, $SE = 0.13$, $p < .02$), such that +/−orthography had a bigger effect on CPF for the LA speaker (*−orth* $M = 3.49$, *+orth* $M = 4.40$) than for the MA speaker (*−orth* $M = 2.97$, *+orth* $M = 3.51$).

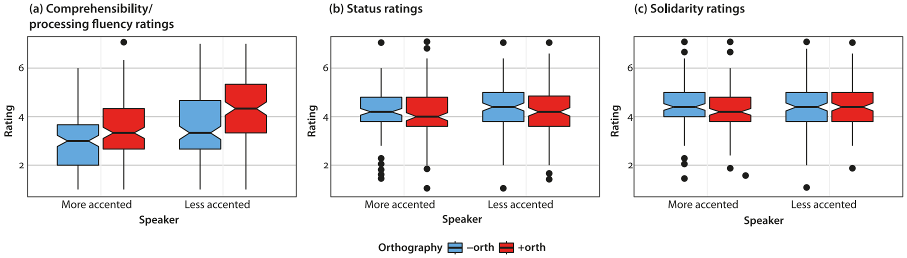


Figure 2. Experiment 1 results
Note. Raw (a) CPF, (b) status, and (c) solidarity ratings. Significant findings: Interaction between speaker and orthography for CPF ratings, and main effect for speaker for status ratings.

2.2.2 *Focal analyses*

Given that both manipulations were successful in influencing CPF, we examined whether they affected listeners' social evaluations. Two separate models were run, one for composite status scores and one for composite solidarity scores as dependent variables. For status (Figure 2b), listeners' ratings significantly differed by speaker ($\beta = 0.15$, $SE = 0.05$, $p < .004$), with the LA speaker rated more highly ($M = 4.32$) than the MA speaker ($M = 4.15$). Orthography was not a significant predictor of status scores ($\beta = -0.11$, $SE = 0.10$, $p < .26$), counter to the processing fluency hypothesis. In fact, numerically, listeners in the *+orth* condition gave more *negative* status ratings ($M = 4.17$) than *−orth* ($M = 4.29$). The control factors passage ($\beta = 0.18$, $SE = 0.05$, $p < .001$, NWS $M = 4.11$, Stella $M = 4.36$) and passage order ($\beta = -0.23$, $SE = 0.10$, $p < .019$, NWSFirst $M = 4.36$, StellaFirst $M = 4.11$), but not speaker order, also significantly affected status ratings. No interactions significantly improved the model.

In terms of solidarity (Figure 2c), there was a significant main effect of passage ($\beta = 0.16$, $SE = 0.06$, $p < .004$; NWS $M = 4.27$, Stella $M = 4.49$), but no other factors or interactions were significant, including speaker and +/−orthography.

2.3 Discussion

In Experiment 1, the processing fluency hypothesis was not supported, with no evidence that the more comprehensible/fluent CPF ratings from listeners in the *+orth* condition predicted more positive status or solidarity ratings. This lack of effect of orthography on status and solidarity aligns with the results of Dragojevic (2020), which failed to find consistent CPF effects. As for speaker accentedness, our findings align with Dragojevic et al. (2017), with the less accented speaker receiving more comprehensible/fluent CPF and more positive status ratings (the lone result predicted by the processing fluency hypothesis), but equivalent solidarity ratings.

3. Experiment 2

Experiment 2 investigated whether the processing fluency hypothesis applies when manipulating CPF within-subjects, and assesses the degree of context-dependence of the relationship between CPF and social evaluation. Blocking trials by +/−orthography allows us to assess not only the effect of orthography (as in Experiment 1), but also whether *changes* in orthography affect the relationship. If listeners' experience of CPF increases with the addition of orthography, are they as likely to increase social evaluation ratings as they would be to decrease such ratings when their experience of CPF decreases? If the relationship between orthography and social evaluation is malleable by context, this suggests limits to the generalizability of the processing fluency hypothesis.

3.1 Materials and methods

3.1.1 *Materials*

The same two speakers from Experiment 1 were used. In order to collect more than one rating per listener per condition (2 speakers × 2 ratings = 4 passages), and because the Wildcat Corpus contains only three reading passages, new passages were constructed by concatenating individual sentences taken from a different portion of the Wildcat Corpus (originally designed for Bradlow & Alexander, 2007). Each of the four new passages contained six thematically related sentences.

The sentences were originally constructed so that the final word was predictable ("A pigeon is a kind of bird") or unpredictable ("He pointed at the animals") from previous context. Though they were not used for that purpose in this study, we balanced predictability across passages, with five high predictability sentences and one low predictability sentence in each passage. Passages were shorter overall than in Experiment 1 (though in this experiment each listener heard two passages per speaker instead of one), and the LA speaker's passages (*M*=17.5 sec) were shorter on average than the MA speaker's (*M*=21 sec). Recordings were amplitude-normalized to yield an approximately equal volume across stimuli.

3.1.2 *Procedure*

The procedure was identical to Experiment 1 except for a few changes. Listeners heard four passages, two from each speaker (the same speaker always produced the same two passages, a concession made to avoid the design becoming too large due to counterbalancing), and thus answered the full set of questions four times. Two between-subjects conditions blocked and counterbalanced orthography order across the four passages: *+orthFirst* (trials 1 & 2: *+orth*, trials 3 & 4: *−orth*); and *−orthFirst* (trials 1 & 2: *−orth*, trials 3 & 4: *+orth*). Within those two conditions, counterbalancing for speaker order and passage order resulted in a total of 8 conditions. Questions were identical to those in Experiment 1. Additionally, for Mechanical Turk participants, one "catch question" per passage was included to ensure participants' attention.

3.1.3 *Participants*

A total of 411 participants took part in Experiment 2. Participants were either undergraduates recruited from the same population as Experiment 1, University of Oregon Human Subjects Pool undergraduates participating online (SP, *N*=257), or were recruited from Amazon's Mechanical Turk (MT, *N*=155), included for generalizability beyond the undergraduate population. Mechanical Turk participants were restricted to IP addresses in the United States, and were paid the equivalent of $ 10/hour for their participation. Three participants were excluded because they learned English after age 5 (1 SP, 2 MT), 31 participants were excluded for not completing the task (9 SP, 22 MT) and 3 MT participants were excluded for not answering catch questions correctly, leaving 376 participants (247 SP, 129 MT). Again, all trials in which participants did not accurately identify that the speaker was from Korea were excluded, resulting in removal of 16.1% of trials (equally distributed across *+orth* and *−orth* conditions), and a total of 10 participants (10 SP, 0 MT). This step left 366 participants (ages 18–69, *M* age=26; 195 female, 167 male, 4 non-binary) with at least one trial for analysis (total trials=1262). All participants were combined for statistical analyses, but participant type (SP or MT) was included as a factor to examine possible differences between populations. As in Experiment 1, the majority of listeners reported having infrequent exposure to Korean-accented English, 65.3% hearing the accent infrequently or very infrequently, 21.0% neither often nor infrequently, and 13.6% often or very often; this information did not improve any model fits and thus is not analyzed further.

3.2 Results

As in Experiment 1, linear mixed effects models were fit to the data with participant as a random intercept. The critical independent variables (speaker, +/−orthography, and orthography order) and other control factors (speaker order and participant type) were included as fixed effects. Due to an a priori interest in the effect of context (orthography order) on the two variables of interest, all modeling stepped down from a three-way interaction between speaker, +/−orthography, and orthography order, keeping only the interactions that improved model fit. Participant type (SP or MT) did not significantly affect any dependent measure as a main effect or interaction, and is not discussed further.

3.2.1 *Manipulation check*

First, we tested whether our manipulations affected CPF (Figure 3a). CPF ratings significantly differed by speaker ($\beta = .35$, $SE = .04$, $p < .001$), with the LA speaker ($M = 4.27$) given higher (more comprehensible/fluent) CPF ratings than the MA speaker ($M = 3.76$). Figure 3d plots model estimates (sjPlot R package, Lüdecke, 2018) for the effect of speaker on CPF. Orthography also significantly affected CPF ($\beta = .21$, $SE = .06$, $p < .001$), with *+orth* rated higher (more comprehensible/fluent, $M = 4.32$) than *−orth* ($M = 3.71$). There was also a significant interaction between +/−orthography and orthography order (Figure 3e); $\beta = .39$, $SE = .09$, $p < .001$), such that +/−orthography had a bigger effect on CPF for *+orthFirst* (*−orth* $M = 3.54$, *+orth* $M = 4.45$) than *−orthFirst* (*−orth* $M = 3.90$, *+orth* $M = 4.23$), but simple effects tests revealed that the effect of orthography on CPF was significant in both condition orders (*−orthFirst*: t = −3.50, $p < .001$; *+orthFirst*: t = −9.72, $p < .001$).

3.2.2 *Focal analyses*

Next we examined whether speaker and orthography affected the central measures of interest, listeners' social evaluations (Figures 3b and 3c). For status, the effect of speaker was significant ($\beta = .07$, $SE = .03$, $p < .027$, Figure 3f), such that the LA speaker ($M = 4.43$) had more positive status ratings than the MA speaker ($M = 4.34$). Orthography was not a significant main effect but instead predicted

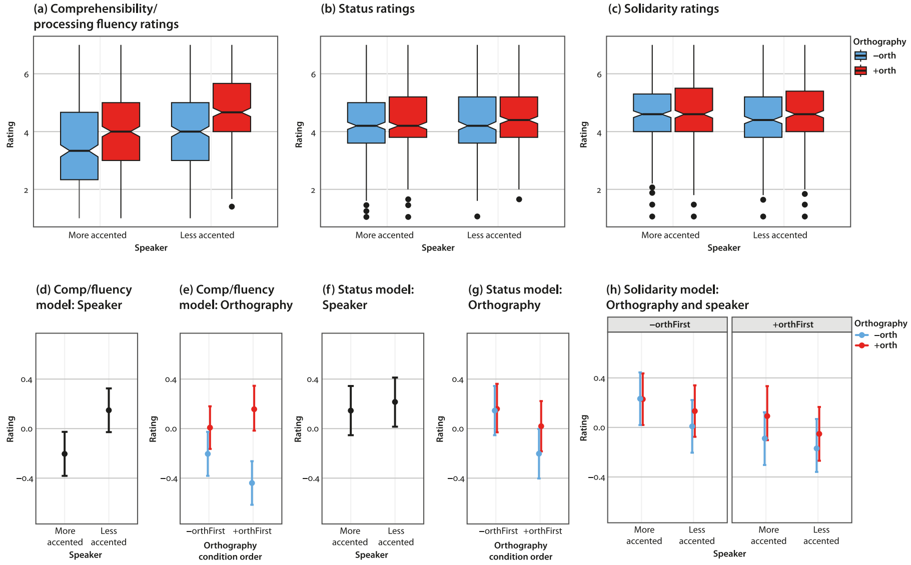


Figure 3. Experiment 2 results

Note. Row 1: raw (a) CPF, (b) status, and (c) solidarity ratings by orthography condition and speaker. Row 2: model estimates of significant main effects: speaker for (d) CPF and (f) status, and interactions: +/−orthography and orthography order for (e) CPF and (g) status; and (h) +/orthography, orthography order, and speaker for solidarity.

status ratings only in interaction with orthography order ($\beta = .20$, $SE = .06$, $p < .002$), where the effect of orthography was bigger for +*orthFirst* listeners, −*orth* $M = 4.08$, +*orth* $M = 4.38$, than −*orthFirst* listeners, −*orth* $M = 4.57$, +*orth* $M = 4.53$). A simple effects test revealed that there was no significant effect of orthography for the −*orthFirst* listeners ($t = -.47$, $p < .64$); rather, status ratings by +/−orthography differed for the +*orthFirst* listeners only ($t = -4.88$, $p < .0001$). Further, there was a significant main effect of orthography order ($\beta = -.35$, $SE = .10$, $p < .001$), where +*orthFirst* listeners ($M = 4.21$) assigned more negative status ratings overall than −*orthFirst* listeners ($M = 4.55$). Figure 3g illustrates +/−orthography's effect; in the +*orthFirst* condition, where listeners began with intermediate status ratings in their +*orth* trials (red), and then downgraded their ratings in −*orth* trials (blue). In contrast, −*orthFirst* listeners started out giving relatively more positive status ratings in −*orth* trials (blue), but did not boost status ratings in the +*orth* trials (red), leading to no difference by +/−orthography.

Results for solidarity ratings showed a more complex set of predictors. Speaker contributed to solidarity scores ($\beta = -.22$, $SE = .06$, $p < .001$), but not as predicted by the processing fluency hypothesis; the MA speaker ($M = 4.62$) received more *positive* solidarity ratings than the LA speaker ($M = 4.45$). Orthography was not a significant main effect, but contributed to solidarity ratings in several interactions. The three-way interaction between speaker, +/−orthography, and orthography order was significant ($\beta = -.23$, $SE = .12$, $p < .043$), as were two of the two-way interactions contained therein. The interaction between +/−orthography and orthography order was significant ($\beta = .20$, $SE = .08$, $p < .012$), where the effect of orthography was greater for *+orthFirst* (*−orth* $M = 4.36$, *+orth* $M = 4.57$), than for *−orthFirst* (*−orth* $M = 4.71$, *+orth* $M = 4.70$). The interaction between speaker and orthography order was also significant ($\beta = .16$, $SE = .08$, $p < .043$), where the condition order mattered more for the MA speaker (*−orthFirst* $M = 4.80$, *+orthFirst* $M = 4.50$), than the LA speaker (*−orthFirst* $M = 4.61$, + *orthFirst* $M = 4.40$). Further, there was a significant main effect of orthography order ($\beta = -.32$, $SE = .10$, $p < .002$, where *−orthFirst* listeners, $M = 4.71$, gave more positive ratings than *+orthFirst* listeners, $M = 4.45$). As is visible in Figure 3h, simple contrasts reveal that the sole case where the effect of orthography was significant was for the MA speaker in the *+orthFirst* condition ($t = -3.46$, $p < .001$), That is, listeners in the *+orthFirst* condition significantly downgraded the more accented speaker on solidarity in *−orth* trials, but no other comparisons reached significance.

In sum, listeners showed similar overall trends when assigning solidarity traits and status traits, with the caveats that for solidarity ratings: the more accented speaker was given more positive ratings than the less accented speaker, and the interaction between +/−orthography and orthography order differed by speaker. Overall, orthography affected status or solidarity ratings only in interaction with context manipulations.

3.3 Discussion

In terms of the effect of speaker accentedness, the LA speaker was assigned a more positive status rating (as in Experiment 1), but the opposite pattern was found for solidarity (though only for *+orthFirst* listeners who also downgraded the MA speaker after orthography was removed). These differences emerged despite the fact that only a minority of listeners correctly identified the number of speakers they heard: 39.6% reported that they heard two speakers read two passages each (compared to the incorrect responses: a different speaker read each passage, 46.0%; or the same speaker read all passages, 14.4%).

In terms of the effect of orthography on status and solidarity ratings, the contextual factor orthography order played a major role, having a larger impact on listeners' behavior than the speaker or orthography manipulations on their own. For both status and solidarity ratings, the orthography manipulation was significant only for *+orthFirst* listeners, and for solidarity only for *+orthFirst* listeners rating the MA speaker. In general, then, this experiment indicates that the effect of orthography was conditional; it did not apply across the board.

We note that had we only examined the first two trials in Experiment 2 (resulting in a between-subjects design as in Experiment 1), +/orthography would have had a null effect for both status and solidarity; differences in social evaluations by orthography were only contributed by later trials, where listeners could make an implicit comparison with already experienced trials. We discuss potential accounts for this finding in the General discussion.

4. General discussion

This paper examined the processing fluency hypothesis, which posits a causal effect of comprehensibility/processing fluency on social evaluation. Two experiments compared listeners' ratings of two L1-Korean English speakers when orthographic representations of their speech were or were not presented concurrently with audio recordings. In Experiment 1, varying orthography across groups, we found that listeners receiving orthography gave more comprehensible/fluent CPF ratings than those with no orthography, but social evaluation ratings did not vary. Experiment 2 varied orthography within groups and found an effect of orthography on social evaluation ratings only in cases where participants received orthography first and then saw no orthography. Taken together, these findings add to the mixed results from previous research investigating the processing fluency hypothesis. Manipulations affecting CPF ratings without regard for social factors (e.g., noise, or orthography in our study) appear to have a fragile effect on social evaluation across studies. And, our findings offer the new insight that the relationship between CPF and social evaluation is highly susceptible to listeners' context-induced experience, as operationalized here by orthography order.

In terms of broader applications of these findings, the assumption that researchers make about the cause of negative social evaluations of L2 learners (Route a vs. b from Figure 1) has implications for which interventions could improve societal outcomes for L2 learners. For example, if researchers believe that negative social evaluations of L2 speakers arise from category-driven stereotypes (as in Route a, in line with the reverse linguistic stereotyping account), interventions geared toward changing listeners' stereotypes are called for in promoting more positive social evaluation. On the other hand, if negative social evaluations are thought to be the result of listeners' difficulties in processing L2 speech (as in Route b, the processing fluency account), then making it easier for listeners to process L2 speech should result in more positive social evaluations. In our study, we in fact performed this latter intervention, by presenting orthography.

We found that reading the speech alongside hearing it made processing easier for listeners, but positive social evaluations of L2 speakers did not uniformly follow. This outcome suggests that processing difficulties are not the major cause of negative social evaluations, and therefore other types of interventions, such as those aiming to affect listeners' stereotypes and ideology, might prove more promising.

Orthography's conditional effect on status and solidarity ratings has several possible accounts, which are not mutually exclusive. First, since orthography affected CPF ratings more for *+orthFirst* listeners, CPF itself may have had more of an effect on social evaluations in the *+orthFirst* conditions (though unlike social evaluations, orthography also affected CPF in *−orthFirst* conditions). This account may partially explain the mixed results in prior studies. The relationship between CPF and social evaluation may be more salient to listeners in cases where they begin to experience more processing difficulty than they expect to, as in *+orthFirst* listeners' later trials, where the previously present orthography was removed. Since CPF manipulations in prior studies have been conducted between-subjects, it is hard to disentangle the role of listener's prior experiences from the CPF manipulations; the manipulation may matter more relative to expectations, which are harder to assess in such a design.

Second, the fact that listeners were more willing to socially downgrade rather than upgrade a speaker they had previously rated may reflect a more general property of social cognition. The asymmetrical relativity of social evaluation in this study echoes earlier social psychological studies finding a negativity bias in reframing effects in social perception: attitudes change more when an initially positive frame switches to negative, as compared with negative to positive reframing (Sparks & Ledgerwood, 2017). For example, it takes less negative information for perceivers to downwardly revise positive first impressions of a person than it takes positive information to upwardly revise negative first impressions (Klein & O'Brien, 2016). Thus, social evaluations may be more generally subject to a negativity bias across contexts.

Major challenges remain in establishing causality in the relationship between CPF and social evaluation. For example, the constructs were measured here by collecting ratings from listeners after playing the speech recordings, but ratings from the same individual are inherently linked and often correlate (Ogden, 2019), making it difficult to establish causation even for conceptually independent constructs. One promising path forward is the use of objective measures of CPF like pupillometry that are orthogonal to subjective ratings, such as the approach taken by Ogden (2019). Pupil dilation has been shown to track listening effort, even in completely intelligible speech (McLaughlin & Van Engen, 2020). Pupillometry has many of the same advantages as measuring processing time as an index of comprehensibility (Munro & Derwing, 1995b), with the added advantage that it is decoupled from listeners' responses and can dynamically track CPF throughout a stimulus. Thus, using the methods employed thus far in the literature (including in the present study), it is still not possible to determine whether CPF *caused* changes in social evaluations. Since the present study experimentally manipulated orthography we can infer that orthography is causally responsible for CPF and social evaluation measures, but not that orthography's effect is due to CPF. It is possible, for example, that CPF and social evaluation correlate because of some yet untested factor.

For instance, attention may play a role. Attention's impact could work in either direction: In the presence of orthography, listeners may not attend as closely to the signal and thus encode fewer indexically signaled traits of the speakers. Alternatively, with orthography listeners may have more resources to allocate to attending to the signal and could therefore encode more deeply. The current experiment was not designed to investigate this question, which we leave open for future work. However, one participant stated in an open-ended comment about the task, "I paid less attention to the speaker when the words were in front of me and had a harder time remembering things about the speaker", consistent with a decreased attention in *+orth* trials explanation. More generally, it may be that when listening to nonnative speech, listeners devote less attention to the bottom-up signal (Lev-Ari & Keysar, 2012). Thus, stimuli receiving low CPF ratings may also receive less listener attention to acoustic details.

However, the degree of attention paid to the signal may itself be contextually modulated. That is, a listener's social classification of a speaker (e.g., in Route a in Figure 1) affects the amount of attention allocated to processing their speech (Sumner et al., 2014). For example, an unfamiliar stigmatized native variety can induce processing costs in word recognition where an equally unfamiliar prestigious variety does not (Sumner & Kataoka, 2013). Therefore, although there has been an understandable effort in earlier research to isolate the role of CPF from the role of social information like stereotypes in assigning social evaluations (Dragojevic & Giles, 2016), the underlying social landscape cannot be ignored even when attempting to investigate "purely" cognitive constructs. For social evaluations, this reminder is especially important. For example, Hall-Lew et al. (2019) present qualitative evidence that in a situation where local knowledge is prized (Scottish heritage tourism) and speech serves as a marker of locality, low intelligibility is equated with *high* credibility; a Scottish-sounding tour guide lends authenticity to a tour of Edinburgh. This observation serves to reinforce that social evaluations are always made within a particular context.

In Experiment 2 we investigated the role of one task-related context, orthography order, but ideological contexts are likely more relevant. The broader social context of any judgment about a speaker is not ideologically neutral, even in research, educational, and assessment settings. In these settings, where expert or naive listeners explicitly rate others' speech, the ideological backdrop is couched in a standard (if not native speaker, see Levis, 2018) norm. Ignoring this backdrop or accepting it as universal makes it dangerous to appeal to *purely* cognitive explanations for social evaluations (see also Hall-Lew et al., 2019), as cognition is socially embedded. In contexts where language learners are authorities, would listeners' experiences of CPF and their social evaluations of those learners show the same relationship (Zuengler & Bent, 1991)? With this in mind, we return to the two potential causal relationships between CPF and social evaluation depicted in Figure 1 and suggest that seeing them as discrete, oppositional pathways may in fact be an artificial division. Future work should keep in mind the relative, context-dependent nature of these constructs, and take into account the embeddings of these relationships within a variety of situational, cognitive, and social contexts.

5. Conclusion

This study investigated the relationship between comprehensibility/processing fluency and social evaluation. Our findings cannot determine causality, but indicate a conditional relationship between CPF and social evaluation, suggesting limits to the processing fluency hypothesis. Further, that CPF and social evaluations were influenced by contextual experience reminds us to consider context in all measures of learner speech. We suggest that better understanding how social evaluation relates to comprehensibility, accentedness, and intelligibility is an important long-term research goal, not only for theory and practice, but also for improving social outcomes for L2 learners.

References

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, 13(3), 219–235. https://doi.org/10.1177/1088868309341564

Bradlow, A. R. & Alexander, J. A. (2007). Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners. *The Journal of the Acoustical Society of America*, 121(4): 2339–2349. https://doi.org/10.1121/1.2642103

Brennan, E. M. & Brennan, J. S. (1981). Accent scaling and language attitudes: Reactions to Mexican American English Speech, *Language and Speech*, 24(3), 207–221. https://doi.org/10.1177/002383098102400301

Bresnahan, M.J., Ohashi, R., Nebashi, R., Liu, W.Y., & Shearman, S.M. (2002). Attitudinal and affective response toward accented English. *Language and Communication*, 22(2), 171–185. https://doi.org/10.1016/S0271-5309(01)00025-8

Cardoso, A., Levon, E., Sharma, D., Watt, D., & Ye, Y. (2019). Inter-speaker variation and the evaluation of British English accents in employment contexts. In Calhoun, S., Escudero, P., Tabain, M., & Warren, P. (Eds.) *Proceedings of the International Congress of Phonetic Sciences* (1615–1619). Melbourne, Australia.

De Meo, A., Vitale, M., Pettorino, M. & Martin, P. (2011). Acoustic-perceptual credibility correlates of news reading by native and Chinese speakers of Italian. In W. Lee & E. Zee (Eds.) *Proceedings of the International Congress of Phonetic Sciences* (1366–1369). Hong Kong.

Derwing, T.M., & Munro, M.J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, 19(1), 1–16. https://doi.org/10.1017/S0272263197001010

Derwing, T.M., & Munro, M.J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching*, 42(4), 476–490. https://doi.org/10.1017/S026144480800551X

Dixon, J.A., & Mahoney, B. (2004). The effect of accent evaluation and evidence on a suspect's perceived guilt and criminality. *The Journal of Social Psychology*, 144(1), 63–73. https://doi.org/10.3200/SOCP.144.1.63-73

Dragojevic, M. (2020). Extending the fluency principle: Factors that increase listeners' processing fluency positively bias their language attitudes. *Communication Monographs*, (87)2, 158–178. https://doi.org/10.1080/03637751.2019.1663543

Dragojevic, M., & Giles, H. (2014). The reference frame effect: An intergroup perspective on language attitudes. *Human Communication Research*, 40(1), 91–111. https://doi.org/10.1111/hcre.12017

Dragojevic, M., & Giles, H. (2016). I don't like you because you're hard to understand: The role of processing fluency in the language attitudes process. *Human Communication Research*, 42(3), 396–420. https://doi.org/10.1111/hcre.12079

Dragojevic, M., Giles, H., Beck, A.C., & Tatum, N.T. (2017). The fluency principle: Why foreign accent strength negatively biases language attitudes. *Communication Monographs*, 84(3), 385–405. https://doi.org/10.1080/03637751.2017.1322213

Giles, H., & Rakić, T. (2014). Language attitudes: Social determinants and consequences of language variation. In T.M. Holtgraves (Ed.), *The Oxford Handbook of Language and Social Psychology*, 11–26. New York, NY: Oxford University Press.

Gluszek, A., & Dovidio, J.F. (2010). The way they speak: A social psychological perspective on the stigma of nonnative accents in communication. *Personality and Social Psychology Review*, 14(2), 214–237. https://doi.org/10.1177/1088868309359288

Hall-Lew, L., Paiva Couceiro, I., & Fairs, A. (2019). Credibility without intelligibility: Implications for hearing vernacular speakers. In R. Blake, & I. Buchstaller (Eds.), *The Routledge Companion to the Work of John R. Rickford* (220–230). New York, NY: Taylor & Francis. https://doi.org/10.4324/9780429427886-23

Hosoda, M., & Stone-Romero, E. (2010). The effects of foreign accents on employment-related decisions. *Journal of Managerial Psychology*, 25(2), 113–132. https://doi.org/10.1108/02683941011019339

Hu, G., & Lindemann, S. (2009). Stereotypes of Cantonese English, apparent native/non-native status, and their effect on non-native English speakers' perception. *Journal of Multilingual and Multicultural Development*, 30(3), 253–269. https://doi.org/10.1080/01434630802651677

Ingvalson, E. M., Lansford, K. L., Federova, V., & Fernandez, G. (2017a). Listeners' attitudes toward accented talkers uniquely predicts accented speech perception. *The Journal of the Acoustical Society of America*, 141(3), 234–238. https://doi.org/10.1121/1.4977583

Ingvalson, E. M., Lansford, K. L., Fedorova, V., & Fernandez, G. (2017b). Cognitive factors as predictors of accented speech perception for younger and older adults. *The Journal of the Acoustical Society of America*, 141(6), 4652–4659. https://doi.org/10.1121/1.4986930

Kang, O., & Rubin, D. L. (2009). Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. *Journal of Language and Social Psychology*, 28(4), 441–456. https://doi.org/10.1177/0261927X09341950

Klein, N., & O'Brien, E. (2016). The tipping point of moral change: When do good and bad acts make good and bad actors? *Social Cognition*, 34(2), 149–166. https://doi.org/10.1521/soco.2016.34.2.149

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software*, 82(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lev-Ari, S., & Keysar, B. (2010). Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of Experimental Social Psychology*, 46(6), 1093–1096. https://doi.org/10.1016/j.jesp.2010.05.025

Lev-Ari, S., & Keysar, B. (2012). Less-detailed representation of non-native language: Why non-native speakers' stories seem more vague. *Discourse Processes*, 49(7), 523–538. https://doi.org/10.1080/0163853X.2012.698493

Levis, J. (2018). *Intelligibility, oral communication, and the teaching of pronunciation*. New York, NY: Cambridge University Press. https://doi.org/10.1017/9781108241564

Lick, D. J., & Johnson, K. L. (2015). The interpersonal consequences of processing ease: Fluency as a metacognitive foundation for prejudice. *Current Directions in Psychological Science*, 24(2), 143–148. https://doi.org/10.1177/0963721414558116

Lindemann, S. (2003). Koreans, Chinese or Indians? Attitudes and ideologies about non-native English speakers in the United States. *Journal of Sociolinguistics*, 7(3), 348–364. https://doi.org/10.1111/1467-9481.00228

Lindemann, S., & Subtirelu, N. (2013). Reliably biased: The role of listener expectation in the perception of second language speech. *Language Learning*, 63(3), 567–594. https://doi.org/10.1111/lang.12014

Lippi-Green, R. (2012). *English with an accent: Language, ideology, and discrimination in the United States* (2nd ed). London, England: Routledge. https://doi.org/10.4324/9780203348802

Lüdecke, D. (2018). *sjPlot: Data Visualization for Statistics in Social Science*. R package version 2.4.1.9000.

McGowan, K. B. (2015). Social expectation improves speech perception in noise. *Language and Speech*, 58(4), 502–521. https://doi.org/10.1177/0023830914565191

McLaughlin, D. J., Baese-Berk, M. M., Bent, T., Borrie, S. A., & Van Engen, K. J. (2018). Coping with adversity: Individual differences in the perception of noisy and accented speech. *Attention, Perception, & Psychophysics*, 80(6), 1559–1570. https://doi.org/10.3758/s13414-018-1537-4

McLaughlin, D.J., & Van Engen, K.J. (2020). Task-evoked pupil response for accurately recognized accented speech. *The Journal of the Acoustical Society of America*, 147(2), 151–156. https://doi.org/10.1121/10.0000718

Munro, M.J. (2018). Dimensions of pronunciation. In O. Kang, R. Thomson, & J. Murphy (Eds.), *The Routledge Handbook of Contemporary English Pronunciation* (413–431). New York: Routledge.

Munro, M.J. & Derwing, T.M. (1995a). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, 45(1), 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Munro, M.J., & Derwing, T.M. (1995b). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language & Speech*, 38(3), 289–306. https://doi.org/10.1177/002383099503800305

Nagle, C., Trofimovich, P., & Bergeron, A. (2019). Toward a dynamic view of second language comprehensibility. *Studies in Second Language Acquisition*, 41(4), 647–672. https://doi.org/10.1017/S0272263119000044

Ogden, D. (2019). Processing fluency, perceptual adaptation, and language attitudes: Does adaptation improve comprehension ease and attitudes toward speakers with non-native accents? [Unpublished doctoral dissertation]. University of Michigan.

Ryan, E.B., Carranza, M.A. and Moffie, R.W. (1977). Reactions toward varying degrees of accentedness in the speech of Spanish-English bilinguals. *Language and Speech*, 20, 267–273. https://doi.org/10.1177/002383097702000308

Ryan, E.B., Giles, H., & Sebastian, R.J. (1982). An integrative perspective for the study of attitudes toward language variation. In E.B. Ryan & H. Giles (Eds.), *Attitudes Towards Language Variation: Social and Applied Contexts* (1–19). London: Edward Arnold.

Souza, A.L., & Markman, A.B. (2013). Foreign accent does not influence cognitive judgments. In *Proceedings of the Annual Meeting of the Cognitive Science Society* 35(35) 1360–1365.

Sparks, J., & Ledgerwood, A. (2017). When good is stickier than bad: Understanding gain/loss asymmetries in sequential framing effects. *Journal of Experimental Psychology: General*, 146(8), 1086–1105. https://doi.org/10.1037/xge0000311

Stocker, L. (2017). The Impact of Foreign Accent on Credibility: An Analysis of Cognitive Statement Ratings in a Swiss Context. *Journal of Psycholinguistic Research*, 46(3), 617–628. https://doi.org/10.1007/s10936-016-9455-x

Sumner, M. and Kataoka, R. (2013). Effects of phonetically-cued talker variation on semantic-encoding. *The Journal of the Acoustical Society of America*, 134(6), EL485–EL491. https://doi.org/10.1121/1.4826151

Sumner, M., Kim, S.K., King, E., & McGowan, K.B. (2014). The socially-weighted encoding of spoken words: A dual-route approach to speech perception. *Frontiers in Psychology*, 4(1015) 1–13. https://doi.org/10.3389/fpsyg.2013.01015

Taylor Reid, K., Trofimovich, P., & O'Brien, M.G. (2019). Social attitudes and speech ratings: Effects of positive and negative bias on multiage listeners' judgments of second language speech. *Studies in Second Language Acquisition*, 41(2), 419–442. https://doi.org/10.1017/S0272263118000244

Tzeng, C.Y., Alexander, J.E., Sidaras, S.K., Nygaard, L.C. (2016). The role of training structure in perceptual learning of accented speech. *Journal of Experimental Psychology: Human Perception and Performance*, 42(11), 1–17.

Van Engen, K. J., Baese-Berk, M. M., Baker, R. E., Choi, A., Kim, M., & Bradlow, A. R. (2010). The Wildcat Corpus of native-and foreign-accented English: Communicative efficiency across conversational dyads with varying language alignment profiles. *Language & Speech*, 53(4), 510–540. https://doi.org/10.1177/0023830910372495

Vaughn, C. (2019). Expectations about the source of a speaker's accent affect accent adaptation. *The Journal of the Acoustical Society of America*, 145(5), 3218–3232. https://doi.org/10.1121/1.5108831

Vaughn, C., & Baese-Berk, M. (2019). Effects of talker order on accent ratings. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences.* Melbourne, Australia.

Zuengler, J., & Bent, B. (1991). Relative knowledge of content domain: An influence on native-non-native conversations. *Applied Linguistics*, 12(4), 397–415. https://doi.org/10.1093/applin/12.4.397

Subject index

Inspired by Murray Munro and Tracey Derwing's 1995 seminal study of intelligibility, comprehensibility, and accentedness, this book revisits the insights of their original research and presents subsequent studies extending this work to new ways of understanding second language speech. By rejecting the nativeness approach upon which previous pronunciation research and teaching were built, Munro and Derwing's paper became the catalyst for a new paradigm of pronunciation and speech research and teaching. For the first time, pronunciation researchers had an empirically-motivated set of dimensions for assessing L2 speech. Results of many subsequent studies showed that the original insights of three partially-independent measures are indispensable to language teaching, language assessment, social evaluations of speech, and pedagogical priorities. This monograph offers 9 diverse chapters by leading researchers, all of which focus on intelligibility and or comprehensibility. This volume is essential reading for anyone interested in up-to-date coverage of L2 pronunciation matters.

Originally published as special issue of *Journal of Second Language Pronunciation* 6:3 (2020)

ISBN 978 90 272 1137 8

9 789027 211378

John Benjamins Publishing Company