14 Second Language Pronunciation Assessment: A Look at the Present and the Future

Pavel Trofimovich and Talia Isaacs

Introduction

Over three decades ago, Michael Canale summarized what he considered to be the challenges facing language assessment in the era of communicative language learning and teaching:

> Just as the shift in emphasis from language form to language use has placed new demands on language teaching, so too has it placed new demands on language testing. Evaluation within a communicative approach must address, for example, new content areas such as sociolinguistic appropriateness rules, new testing formats to permit and encourage creative, open-ended language use, new test administration procedures to emphasize interpersonal interaction in authentic situations, and new scoring procedures of a manual and judgemental nature. (Canale, 1984: 79)

Applied to second language (L2) pronunciation assessment, this description remains highly relevant today, raising a number of important issues, such as: broadening the scope of pronunciation assessment beyond the focus on a single aspect of pronunciation (e.g. segmental accuracy) or a single standard (e.g. absence of a discernible nonnative accent); targeting pronunciation assessment for various interlocutors in interactive settings, for instance, outside a typical focus on academic performance by students in Western societies; as well as developing and fine-tuning novel assessment instruments and procedures. Above all, Canale's description aptly summarizes an ongoing quest in language assessment to capture the authenticity and interactiveness of

language use (e.g. Bachman, 1990; Bachman & Palmer, 2010). The contributions to this edited volume address some of the challenges identified by Canale in innovative ways. Before summarizing these contributions, we hasten to add that no edited volume, including this one, can provide an exhaustive overview of all possible issues in L2 pronunciation assessment; most chapters in this volume are focused on testing or informal evaluative judgements of speech in real-world settings and not on classroom-based assessment, including diagnostic assessment or feedback on test takers' performance. However, the range of topics, the variety of research methodologies and paradigms, and the scope of implications featured here make this volume a timely addition to the growing area of L2 pronunciation assessment.

Current Trends

A focus on intelligibility

According to Levis (2005: 370) and echoed in **Harding**'s chapter, teaching and, by extension, assessing L2 pronunciation can be characterized as the tension between two 'contradictory principles'. The nativeness principle holds that nativelike, unaccented pronunciation is both a chief goal of pronunciation learning and a standard for pronunciation assessment. By contrast, the intelligibility principle posits that the primary goal of pronunciation learning is for learners to be understood by their interlocutors, with the consequence that intelligibility, rather than nativeness, emerges as an appropriate assessment criterion. The research findings are clear: a noticeable or even strong nonnative accent does not always involve a lack of understanding (Derwing & Munro, 2015).

While most applied linguists would agree that intelligibility, rather than a native accent, should be considered as the appropriate target of pronunciation teaching and learning, the uptake and implementation of the intelligibility construct in language assessment have seen multiple shortcomings. One example of such limitations comes from **Harding**'s qualitative analyses of focus group discussions targeting raters' experience using the Common European Framework of Reference (CEFR) Phonological control scale to rate L2 pronunciation (Council of Europe, 2001). One of the most telling outcomes of this study is that raters believe the scale to be skewed in its treatment of accented versus understandable speech and also to include erratic descriptions of pronunciation features across scale levels. For instance, while lower levels of the scale make reference to speakers' accent, its higher levels refer to intelligibility as a criterion or exclude reference to either construct altogether. Harding reports that, in operational uses of the scale, raters appear to be 'filling in' gaps in scale descriptors, attempting to balance a focus on accent with the perceived need for speakers to be intelligible. This

is an important finding as it not only highlights possible weaknesses of the CEFR phonological control scale but also illustrates how a scale can be developed and refined through consultations with its end-users (raters). Above all, Harding's research raises important questions about the usability, practicality, and – ultimately – validity of scale-based assessments of L2 pronunciation.

In another chapter, featuring a prominent focus on pronunciation constructs related to listener understanding of L2 speech, **Ballard and Winke** investigate the interplay between speakers' accent and comprehensibility (degree of listeners' understanding) and their acceptability as an ESL teacher, focusing on nonnative listeners. They show that nonnative listeners can easily distinguish between accented speakers and those who sound unaccented. Despite this, nonnative listeners do not seem to readily label accented speakers as unacceptable teachers. Instead, listeners associate speakers' acceptability as a teacher with their perceptions of these speakers' comprehensibility. This finding is important in that it confirms that raters' decisions with real-life consequences might depend more strongly on how easily L2 speech is understood rather than on how unaccented it sounds, echoing previous work by Derwing and Munro (2009), which showed a similar result for nonnative English speaking engineers in an English-medium workplace setting.

A focus on language

If listener understanding, whether termed intelligibility or comprehensibility, is an important assessment criterion, then identifying linguistic barriers to communication can help researchers and teachers isolate pronunciation elements to target in teaching and assessment. A vibrant area of research is the relationship between L2 speakers' comprehensibility, frequently operationalized as the extent of listeners' perceived ease or difficulty of understanding L2 speech as measured using a Likert-type rating scale, as in the **Ballard and Winke** study, and linguistic features that characterize their speech, with the goal of helping teachers, learners and language testers to isolate and then focus on features that are most important for listeners' understanding.

Illustrating this line of research, the chapter by **Saito, Trofimovich, Isaacs and Webb** examines a range of linguistic dimensions which could contribute to listeners' judgements of comprehensibility and which, by extension, could elucidate the properties of the speech that listeners (raters) tend to take into account in their scoring, hence enhancing our understanding of the L2 comprehensibility construct. This study is innovative in that it broadens the scope of linguistic factors linked to comprehensibility to include various lexical dimensions of L2 speech, including lexical polysemy, diversity and appropriateness, as well as morphological accuracy. Comprehensibility emerges as a complex construct, spanning various dimensions of speech,

with the consequence that the teaching and assessment of comprehensible L2 speech should consider not only pronunciation and fluency aspects of speech but also its lexical content, such as the use of appropriate and diverse vocabulary. The extent to which lexical features are sensitive to differences in L2 learners' comprehensibility scores across task type (Crowther *et al.*, 2015) also requires further exploration.

In another study focusing on language, **Galaczi, Post, Li, Barker and Schmidt** target rhythm, one dimension of speech prosody, investigating the extent to which several measures of rhythm could distinguish L2 pronunciation levels for learners from different language backgrounds across the CEFR language proficiency scale (Council of Europe, 2001). This study is a welcome contribution to research on L2 pronunciation learning and assessment because it shows that micro-level measures of rhythm, such as speech rate and duration differences between stressed and unstressed syllables, while being useful overall, might not be precise enough to distinguish fine-grained prosodic differences between adjacent levels of the CEFR scale. This finding adds to a growing body of research in language assessment (Isaacs *et al.*, 2015; Iwashita *et al.*, 2008; Kang, 2013; Kang & Wang, 2014) suggesting that various linguistic measures of L2 pronunciation often fail to distinguish between adjacent levels in multi-level pronunciation scales. And because such scales often rely on listener judgements, this finding raises a related question of how well listeners distinguish fine-grained linguistic differences, especially when using scales featuring seven or more levels.

A focus on pronunciation standards

One of the core issues in L2 assessment concerns the standards or criteria by which various aspects of L2 speech are assessed. As discussed previously, intelligible L2 pronunciation – as distinct from L2 pronunciation that sounds unaccented – is typically considered to be an appropriate reference for both teaching and assessment because it reflects what is important for communication, that is, speakers' ability to be understood by interlocutors (Derwing & Munro, 2015). Nevertheless, for many language learners and teachers, what sounds like native and accent-free pronunciation remains an important teaching and learning goal (Scales *et al.*, 2006; Subtirelu, 2013; Young & Walsh, 2010).

Several chapters in this volume focus on the issue of appropriate standards and norms for L2 pronunciation assessment. In a delightful chapter, which reads as an armchair conversation with the author, **Davies** problematizes the concept of the native speaker, with reference to the assessment of L2 pronunciation, touching upon such topics as a standard language, accent prestige, and discrimination based on accent. An insightful chapter by **Lindemann** takes these ideas further, discussing the highly variable and therefore elusive nature of 'standard' pronunciation by native speakers.

Lindemann convincingly argues that classifying nonnative speech as being 'standard' or not is highly problematic, at least in part because of listeners' expectations about L2 speech and their often biased perceptions of it (e.g. Kang & Rubin, 2009). She concludes that a deficit-based approach to the teaching and assessment of L2 pronunciation – one based on defining specific speech patterns in terms of 'errors' or deviations from what is expected in a standard norm – is indefensible, calling for language testers to incorporate the construct of the listener into assessment instruments while also trying to minimize any potential listener-based biases.

In two related chapters, both Sewell and Kennedy *et al.* discuss lingua-franca intelligibility as a criterion for L2 pronunciation assessment in situations when one or more interlocutors from different linguistic and cultural backgrounds share a common language. **Sewell** conceptualizes lingua franca intelligibility within a broad functionalist view of language, implying that the linguistic elements most relevant to intelligibility are those that tend to carry the most information in communication (e.g. consonant contrasts tend to do more 'work' in communication, compared to vowel contrasts). He illustrates this approach to intelligibility using the case of English in Hong Kong, arguing for a teaching and assessment criterion that is rooted in intelligibility but informed by local, contextual considerations specific to sociocultural realities of language use. To cite Sewell, '[t]he lingua franca approach acknowledges that the local is global, and vice versa'. In a conceptually related chapter, **Kennedy, Blanchet and Guénette** rely on verbal reports to understand teacher-raters' judgements of L2 speech in the context of using French as a lingua franca in Quebec, Canada. They conclude that teacher-raters show considerable variability in the extent to which they place importance on mutual understanding in lingua franca interactions while evaluating their students' pronunciation. Kennedy *et al.* speculate that individual differences across teachers in their formal training in phonetics and phonology, their teaching experience and their own language learning histories might explain their preference for native speaker versus lingua franca models in evaluating their learners' L2 pronunciations. These researchers conclude with a call for more research into teachers' beliefs about language and communication, so that classroom assessments and pedagogical decisions can be understood in the context of teacher cognitions (e.g. Baker, 2014).

A focus on other L2 skills

Three contributions to the current volume illustrate that the assessment of L2 pronunciation has much to learn from the expertise in assessment of other language skills and components. In a chapter focusing on speech fluency, **Browne and Fulcher** eloquently argue for the importance of considering listeners' and raters' familiarity with L2 speech in assessment of L2 pronunciation, including intelligibility (operationalized through a gap-fill

task) and speech delivery (fluency). Through the use of Rasch analyses, which allow for a simultaneous treatment of both raters' and speakers' performances on the same arithmetic (logit) scale, they show that a measure of L2 intelligibility and a scored measure of speech delivery based on a five-point TOEFL iBT scale (Educational Testing Service, 2009) predictably vary as a function of rater familiarity with L2 speech. These findings reinforce the idea that various constructs subsumed by the umbrella term 'L2 pronunciation', including speech delivery (fluency) and intelligibility, are not simply tied to speakers' performance but also reflect specific characteristics of individual listeners. The study also brings to light the issue that ideally in L2 pronunciation research, listener familiarity effects, when not directly the source of investigation, should be controlled for, although this is difficult to implement in practice. One implication for high-stakes testing settings could be that accredited examiners should be screened for factors such as their degree of familiarity with the accented speech of the test takers (Winke *et al.*, 2013), although it is unclear how this could be put into practice in contexts where test takers from numerous language backgrounds are being assessed.

Working in the field of L2 writing, **Knoch** provides a comprehensive 'roadmap' for various issues in assessing L2 writing, including the development and validation of rating scales, effects of raters and tasks on assessment outcomes, and applications for classroom-based assessment. Knoch's summary is valuable; it not only offers a wealth of evidence-based information from a skill that has benefited from a larger volume of language assessment research, pioneering many of the advancements in, for example, rater training (e.g. Weigle, 1998), but it also highlights current gaps in the assessment of L2 pronunciation. This includes a paucity of research on the development and validation of L2 pronunciation rating scales with an adequately operationalized construct, the need for more research-based evidence for task and listener effects, and the dearth of research into classroom-based pronunciation assessment and self-assessment, as well as interactive and paired assessments of L2 pronunciation.

In a chapter focusing on assessment of L2 listening, **Wagner and Toth** critically examine the extent to which authentic and simplified (scripted) listening comprehension materials are appropriate as assessment materials. A survey of L2 test takers who took either authentic or scripted listening comprehension materials clearly shows that L2 users are aware of important differences across these recorded stimuli, for example, rating scripted materials lower in authenticity and naturalness and being aware that scripted materials include clearer pronunciation and fewer hesitation markers. Wagner and Toth persuasively argue for the use of testing materials that illustrate authentic, natural and representative uses of real-world spoken language if the goal of teaching, learning and assessment is for learners to comprehend authentic L2 speech. This research reminds L2 teachers, researchers and test developers

to consider the issues of authenticity when designing and validating L2 listening and pronunciation tasks.

A focus on individual differences

Research on L2 development of various language skills clearly shows that there are often pronounced differences across individual learners in rates and outcomes of L2 learning (DeKeyser, 2012). L2 pronunciation learning is no exception. For instance, the learning of various linguistic dimensions of L2 speech has been linked to learners' age (Abrahamsson & Hyltenstam, 2009), the quantity and quality of their contact with the L2 (Moyer, 2011), their motivation and cultural sensitivity (Alvord & Christiansen, 2012; Baker-Smemoe *et al.*, 2014; Hardison, 2014), and their willingness to communicate (Baran-Łucarz, 2014; Derwing *et al.*, 2008). These findings clearly argue against a 'one-size-fits-all' approach to pronunciation teaching and assessment, suggesting that different learners can respond differently to the same testing materials and procedures, and that different materials and procedures might be necessary for assessment of diverse populations of learners. In a novel study, **Mora and Darcy** focus on these issues by investigating the relationship between three cognitive variables (attention control, phonological short-term memory, and inhibitory control) and L2 learners' performance on several acoustic and rated measures of their L2 pronunciation. More importantly, the participants in this study are two groups of English language learners – monolingual speakers (monolingual Spanish speakers) and functionally bilingual language users (Spanish-Catalan bilinguals). Mora and Darcy report a complex set of findings, showing links between learners' attention control and phonological memory and their English vowel production, but only for the group of monolingual Spanish learners of English. The researchers speculate that individual differences in L2 users' cognitive capacities can influence how specific learner groups perform in particular assessment tasks and with particular types of assessment materials, calling for more investigations into individual learner differences to better understand contributors to the variability of learners' L2 pronunciation performance.

Future Directions

To go back to Canale's quote from 30 years ago, it is fair to say that language researchers and assessment specialists have made some as yet limited empirical inroads into the assessment of L2 pronunciation, enhancing our understanding of the constructs under investigation and developing and validating novel assessment procedures. A case in point is the recent launch of fully automated speaking tests into the competitive market of standardized language testing products, including Person's Versant tests, Pearson's speaking

component of the PTE Academic (Bernstein *et al.*, 2010) and Educational Testing Service's TOEFL iBT patented automatic speech recognition technology used with the TOEFL iBT practice speaking test, SpeechRater (Zechner *et al.*, 2009). These instruments, the first two of which tend to be used for high-stakes purposes (e.g. a language proficiency certification test for pilots), are scored using automated speech recognition algorithms optimized to predict human scoring using acoustic and temporal correlates of auditory pronunciation measures in addition to machine scored measures of other linguistic phenomena. Concerns within the assessment community have been raised about automated assessments of speech due to the machine scoring system's ability, as yet, only to cope with highly predictable L2 speaking tasks (e.g. Chun, 2008), as opposed to discourse-level extemporaneous speaking tasks that elicit more varied interactional patterns (see Isaacs, 2016). Technology is rapidly improving. However, speech recognition programmers need to be steered away from targeting accent reduction by modelling acoustic phenomena that are easy for the machine to score and, instead, prioritize the linguistic factors that are most consequential for intelligibility.

Despite these and similar technological advances and developments in conceptual thinking, there is ample room for future research to enhance our understanding of the processes and outcomes of pronunciation testing. At a practical level, research into the assessment of pronunciation in languages other than English is virtually non-existent (for a rare exception, see Kennedy *et al.*, this volume), and assessment research targeting multilingual lingua franca L2 users in non-Western, non-academic contexts is lacking. Also limited is research targeting the assessment of sociolinguistic and pragmatic functions of L2 pronunciation, and research incorporating nonnative pronunciation models and standards in assessments. With respect to practical implications of assessment research, in a climate where assessment for learning, formative assessment, learning-oriented assessment and dynamic assessment (in contrast to large-scale testing) is gaining currency in promoting teaching and learning (Turner & Purpura, 2016), it would similarly be important to expand research on classroom-based assessment, including the instructional effectiveness of incidental form-focused instruction (i.e. corrective feedback) on L2 pronunciation development (e.g. Lee & Lyster, 2016; Saito & Lyster, 2012). In addition, the ground is fertile to build on preliminary work regarding learners' self-assessment of pronunciation (Dlaska & Krekeler, 2008; Lappin-Fortin & Rye, 2014), including helping learners calibrate their perceptions to those of their interlocutors, thus minimizing distorted perceptions of their speech (Trofimovich *et al.*, 2016).

At the conceptual level, Canale's (1984: 79) call for new testing instruments and procedures involving 'interpersonal interaction in authentic situations' has largely not been answered, emphasizing the need for more research into interactional paired and group assessments involving an L2 pronunciation component. There has been some preliminary research in this area in recent

years using the Cambridge interactional (collaborative) test tasks in research settings (Isaacs, 2013; Jaiyote, 2015), although future research needs to go further in investigating pronunciation features that account for communication breakdowns specifically and discrepancies in the extent to which interlocutors report understanding one another. Last but not least, more theorizing is needed targeting possible models or theories that can serve as conceptual bases for the assessment of L2 pronunciation. For instance, as Isaacs (2014) argues, models of communicative competence (Bachman, 1990; Bachman & Palmer, 1996) are insufficiently nuanced to capture all of the complexities of pronunciation, particularly in relation to pronunciation perception, production, and (where applicable) orthographic effects (see also Fulcher, 2003). There is a further need to clarify the role of holistic pronunciation-related constructs such as intelligibility in relation to more discrete L2 speech measures and, if possible, to listener/rater/interlocutor variables. Therefore, more theory building is required to understand the nature of the phenomena being targeted through assessment and, specifically, to better understand major global constructs in L2 pronunciation so they can be better operationalized in assessment instruments (Isaacs & Trofimovich, 2012; see Foote & Trofimovich, submitted, for a discussion of various theoretical frameworks of L2 pronunciation learning).

We conclude this chapter (and in fact the entire volume) with a list of possible issues and questions that we consider to be important for future research into L2 pronunciation assessment. Clearly, this list is non-exhaustive, yet in our opinion it identifies several priority research axes which, if followed, have the greatest potential for enhancing both the breadth and depth of our understanding of L2 pronunciation assessment.

- How do different stakeholders perceive assessments of pronunciation in formal and informal contexts? In what ways can technology be used to validate listener perceptions of linguistic phenomena?
- What is the effect of individual differences in listeners' cognitive or attitudinal variables on listeners' (raters') judgements of L2 pronunciation and on speakers' communicative success in real-world settings?
- How can sources of construct-irrelevant variance related to listener background variables (e.g. accent familiarity effects) be mitigated in high-stakes assessments of L2 speech? What are the implications for rater screening and training and for mitigating sources of bias?
- Which pronunciation features should be prioritized in L2 pronunciation instruction and assessment? How can these features feed into the development of valid speaking assessment instruments?
- How can measures of pronunciation and fluency normally used for individual learners in lab-based research settings be adapted for use in naturalistic settings, including in conversational interactions with interlocutors? Similarly, how can stimuli used in lab-based settings be adapted to generate more authentic testing prompts (e.g. Jones, 2015)?

- In light of the current debate on the native speaker standard and the coexistence of multiple varieties of English, what is the appropriate standard or language varieties that learners should be exposed to for listening tests, including audio prompts for integrated test tasks (e.g. Ockey & French, 2014)? For example, could Cook's (1992, 2012) construct of multicompetence form the basis of a target language assessment standard that draws on descriptions of proficient multicompetent learners or test takers rather than native speakers (e.g. Brown, 2013)?
- If intelligibility is valued as an assessment criterion by the applied linguistics community, then how can intelligibility feed into models of communicative competence (Canale & Swain, 1980) or communicative language ability (Bachman, 1990)? Should intelligibility be instructed and assessed in conjunction with pragmatic competence, focusing on utterances that are not only clearly understood, but are also pragmatically appropriate in the context of language use (e.g. Yates, 2014)?
- How can findings on form-focused instruction in L2 learning and teaching, on the instructional effectiveness of pronunciation interventions, including corrective feedback, and longitudinal studies on learner pronunciation development over time, inform formative assessment practices, particularly in classroom settings?
- How can we move beyond Lado (1961), taking into account technological advancements and the latest developments in research and pedagogy, to bring pronunciation assessment out of its time warp and integrate it into mainstream assessment research and practice?
- To parallel calls to foster language educators' assessment literacy (e.g. Fulcher, 2012; Taylor, 2009), how can we improve experienced teachers' and assessment researchers' and practitioners' pronunciation literacy, making it more mainstream and accessible?

References

Abrahamsson, N. and Hyltenstam, K. (2009) Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Language Learning* 59, 249–306.

Alvord, S.M. and Christiansen, D.E. (2012) Factors influencing the acquisition of Spanish voiced stop spirantization during an extended stay abroad. *Studies in Hispanic and Lusophone Linguistics* 5, 239–276.

Bachman, L.F. (1990) *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

Bachman, L.F. and Palmer, A.S. (1996) *Language Testing in Practice: Designing and Developing Useful Language Tests.* Oxford: Oxford University Press.

Bachman, L. and Palmer, A (2010) *Language Assessment in Practice.* Oxford: Oxford University Press.

Baker, A. (2014) Exploring teachers' knowledge of second language pronunciation techniques: Teacher cognitions, observed classroom practices, and student perceptions. *TESOL Quarterly* 48, 136–163.

Baker-Smemoe, W., Dewey, D.P., Bown, J. and Martinsen, R.A. (2014) Variables affecting L2 gains during study abroad. *Foreign Language Annals* 47, 464–486.
Baran-Łucarz, M. (2014) The link between pronunciation anxiety and willingness to communicate in the foreign-language classroom: The Polish EFL context. *Canadian Modern Language Review* 70, 445–473.
Bernstein, J., Van Moere, A. and Cheng, J. (2010) Validating automated speaking tests. *Language Testing* 27, 355–377.
Brown, A. (2013) Multicompetence and second language assessment. *Language Assessment Quarterly*, 10, 219–235.
Canale, M. (1984) Testing in a communicative approach. In G.A. Jarvis (ed.) *The Challenge for Excellence in Foreign Language Education* (pp. 79–92). Middlebury, VT: Northeast Conference for the Teaching of Foreign Languages.
Canale, M. and Swain, M. (1980) Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics* 1, 1–57.
Chun, C.W. (2008) Comments on 'evaluation of the usefulness of the *Versant for English* test: A response': The author responds. *Language Assessment Quarterly* 5 (2), 168–172.
Cook, V.J. (1992) Evidence for multicompetence. *Language Learning* 42 (4), 557–591.
Cook, V. (2012) Multi-competence. In C.A. Chapelle (ed.) *The Encyclopedia of Applied Linguistics* (pp. 3768–3774). Oxford: Wiley-Blackwell.
Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.
Crowther, D., Trofimovich, P., Isaacs, T. and Saito, K. (2015) Does a speaking task affect second language comprehensibility? *Modern Language Journal* 99, 80–95.
DeKeyser, R. (2012) Interactions between individual differences, treatments, and structures in SLA. *Language Learning* 62, 189–200.
Derwing, T.M. and Munro, M.J. (2009) Comprehensibility as a factor in listener interaction preferences: Implications for the workplace. *Canadian Modern Language Review* 66, 181–202.
Derwing, T.M. and Munro, M.J. (2015) *Pronunciation Fundamentals: Evidence-Based Perspectives for L2 Teaching and Research*. Amsterdam: John Benjamins.
Derwing, T.M., Munro, M.J. and Thomson, R.I. (2008) A longitudinal study of ESL learners' fluency and comprehensibility development. *Applied Linguistics* 29, 359–380.
Dlaska, A. and Krekeler, C. (2008) Self-assessment of pronunciation. *System* 36, 506–516.
Educational Testing Service (2009) *The Official Guide to the TOEFL Test* (3rd edn). New York: McGraw-Hill.
Foote, J.A. and Trofimovich, P. (in press) Second language pronunciation learning: An overview of theoretical perspectives. In O. Kang, R.I. Thomson and J. Murphy (eds) *The Routledge Handbook of Contemporary English Pronunciation*. London: Routledge.
Fulcher, G. (2003) *Testing Second Language Speaking*. London: Pearson.
Fulcher, G. (2012) Assessment literacy for the language classroom. *Language Assessment Quarterly* 9 (2), 113–132.
Hardison, D.M. (2014) Changes in second-language learners' oral skills and socio-affective profiles following study abroad: A mixed-methods approach. *Canadian Modern Language Review* 40, 415–444.
Isaacs, T. (2013) International engineering graduate students' interactional patterns on a paired speaking test: Interlocutors' perspectives. In K. McDonough and A. Mackey (eds) *Second Language Interaction in Diverse Educational Settings* (pp. 227–246). Amsterdam: John Benjamins.
Isaacs, T. (2014) Assessing pronunciation. In A.J. Kunnan (ed.) *The Companion to Language Assessment* (pp. 140–155). Hoboken, NJ: Wiley-Blackwell.

Isaacs, T. (2016) Assessing speaking. In D. Tsagari and J. Banerjee (eds) *Handbook of Second Language Assessment* (pp. 131–146). Berlin: DeGruyter Mouton.

Isaacs, T. and Trofimovich, P. (2012) 'Deconstructing' comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition* 34, 475–505.

Isaacs, T., Trofimovich, P., Yu, G. and 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* 4, 1–48.

Iwashita, N., Brown, A., Mcnamara, T. and O'Hagan, S. (2008) Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics* 29 (1), 24–49.

Jaiyote, S. (2015) The relationship between test-takers' L1, their listening proficiency and their performance in pairs. A*RAGs Research Reports Online, AR-A/2015/2*. Manchester: British Council.

Jones, J. (2015) Exploring open consonantal environments for testing vowel perception. Unpublished Master's thesis, University of Melbourne.

Kang, O. (2013) Linguistic analysis of speaking features distinguishing general English exams at CEFR levels. *Research Notes* 52, 40–48.

Kang, O. and Rubin, D.L. (2009) Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. *Journal of Language and Social Psychology* 28, 441–456.

Kang, O. and Wang, L. (2014) Impact of different task types on candidates' speaking performances and interactive features that distinguish between CEFR levels. *Research Notes* 57, 40–49.

Lado, R. (1961) *Language Testing: The Construction and Use of Foreign Language Tests.* London: Longman.

Lappin-Fortin, K. and Rye, B.J. (2014) The use of pre-/posttest and self-assessment tools in a French pronunciation course. *Foreign Language Annals* 47 (2), 300–320.

Lee, A.H. and Lyster, R. (2016) Effects of different types of corrective feedback on receptive skills in a second language: A speech perception training study. *Language Learning*. Published online. doi:10.1111/lang.12167.

Levis, J.M. (2005) Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly* 39, 369–377.

Moyer, A. (2011) An investigation of experience in L2 phonology. *Canadian Modern Language Review* 67, 191–216.

Ockey, G. and French, R. (2014) From one to multiple accents on a test of L2 listening comprehension. *Applied Linguistics*. Published online. doi:10.1093/applin/amu060.

Saito, K. and Lyster, R. (2012) Effects of form-focused instruction and corrective feedback on L2 pronunciation development of /ɹ/ by Japanese learners of English. *Language Learning* 62, 595–633.

Scales, J., Wennerstrom, A., Richard, D. and Wu, S.H. (2006) Language learners' perceptions of accent. *TESOL Quarterly* 40, 715–738.

Subtirelu, N. (2013) What (do) learners want (?): A re-examination of the issue of learner preferences regarding the use of 'native' speaker norms in English language teaching. *Language Awareness* 22, 270–291.

Taylor, L. (2009) Developing assessment literacy. *Annual Review of Applied Linguistics*, 29, 21–36.

Trofimovich, P., Isaacs, T., Kennedy, S., Saito, K. and Crowther, D. (2016) Flawed self-assessment: Investigating self- and other-perception of second language speech. *Bilingualism: Language and Cognition* 19, 122–140.

Turner, C.E. and Purpura, J.E. (2016) Learning-oriented assessment in the classroom. In D. Tsagari and J. Banerjee (eds) *Handbook of Second Language Assessment* (pp. 255–274). Berlin: DeGruyter Mouton.

Weigle, S.C. (1998) Using FACETS to model rater training effects. *Language Testing* 15, 263–287.

Winke, P., Gass, S. and Myford, C. (2013) Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing* 30, 231–252.

Yates, L. (2014) Learning how to speak: Pronunciation, pragmatics and practicalities in the classroom and beyond. *Language Teaching*. Published online. doi.org/10.1017/S0261444814000238.

Young, T.J. and Walsh, S. (2010) Which English? Whose English? An investigation of 'non-native' teachers' beliefs about target varieties. *Language, Culture, and Curriculum* 23, 123–137.

Zechner, K., Higgins, D., Xi, X. and Williamson, D.M. (2009) Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication* 51, 883–895.