Research timeline: Pronunciation assessment

Authors and affiliation and emails:

Talia Isaacs, UCL Centre for Applied Linguistics, UCL Institute of Education, University College London, UK

talia.isaacs@ucl.ac.uk

Luke Harding, Department of Linguistics and English Language, Lancaster University, Lancaster, UK LA1 4YL

l.harding@lancaster.ac.uk

Biodata:

Talia Isaacs is a Senior Lecturer in Applied Linguistics and TESOL at the UCL Centre for Applied Linguistics, UCL Institute of Education, University College London. Her research examines sources of variability in listeners' judgments of speech, including mapping the factors promoting/impeding efficient oral communication in rating scale descriptors. She has taught a range of applied linguistics courses, including in second language acquisition, language assessment, pedagogy and curriculum, oral communication, and research methodology. She currently serves on the Executive Board of the International Language Testing Association (Member-at-Large) and on the Editorial Boards of the *Journal of Second Language Pronunciation*, *Language Assessment Quarterly*, and *Language Testing*.

Luke Harding is a Senior Lecturer in the Department of Linguistics and English Language at Lancaster University. His research is mainly in the area of language testing, specifically listening assessment, pronunciation and intelligibility, and the challenges of World Englishes and English as a Lingua Franca for language assessment. He regularly teaches on Lancaster's MA in Language Testing on courses including Issues in Language Testing, and Statistical

Analyses for Language Testing. He is the test reviews editor for the journal *Language Testing* and is on the editorial boards of *Language Assessment Quarterly* and the *Journal of Second Language Pronunciation*.

Introduction

After an extended period of being on the periphery, numerous advancements in the field of second language (L2) pronunciation over the past decade have led to increased activity and visibility for this subfield within applied linguistics research. As Derwing (2010) underscored in her 2009 plenary at the first annual *Pronunciation in Second Language Learning and Teaching* (*PSLLT*) conference, a record number of graduate students researching L2 pronunciation and subsequently launching into academic positions at international universities assures L2 pronunciation a bright future in research and teacher training. Other indicators of momentum include the focus of a *Language Teaching* timeline on the topic of pronunciation (Munro & Derwing 2011), the appearance of multiple encyclopaedia volumes or handbooks of pronunciation (e.g., Levis & Munro 2013; Reed & Levis 2015), and the establishment of the specialised *Journal of Second Language Pronunciation* in 2015, which constitutes a milestone in the professionalization of the field and 'an essential step toward a disciplinary identity' (Levis 2015, p. 1).

These positive developments notwithstanding, the vast majority of renewed applied pronunciation research activity has been undertaken by researchers in the fields of Second Language Acquisition (SLA), language pedagogy, sociolinguistics, and psycholinguistics. The language assessment community has been slower in its uptake of interest in pronunciation, with few advocates drawing attention to its exclusion from the collective research agenda or underscoring its marginalization as an assessment criterion in L2 speaking tests until recently (e.g., Harding 2013; Purpura 2016). Pronunciation remains under-conceptualized in models of communicative competence/communicative language ability

(Isaacs 2014) and typically receives minimal coverage in standard texts, such as Luoma's (2004) *Assessing speaking* from the Cambridge Language Assessment series. Although there is a dedicated book on assessing grammar and vocabulary in that series, there is none on assessing pronunciation or pragmatics. The treatment of pronunciation in Fulcher's *Language Teaching* timeline on assessing L2 speaking is indicative, in that it is singled out as the only area relevant to the L2 speaking construct that he was 'not able to cover' (2015, p. 201).

However, there are signs suggesting that pronunciation is also beginning to emerge as an important research area in language assessment. For example, whereas only two pronunciation-focused articles were published in the first 25 years of publication of the longest-standing language assessment journal, *Language Testing* (1984–2009), at least one such article per year has appeared in the years since (2010–). Assessment issues have recently been featured in major events on pronunciation teaching and learning (e.g., 2012 *PSLLT* invited roundtable on pronunciation assessment), while pronunciation has been featured in assessment-oriented discussions (e.g., 2013 *Cambridge Centenary Speaking Symposium*, which will feed into a special issue of *Language Assessment Quarterly*; Lim & Galaczi forthcoming). A general shift in attention in language assessment research towards pronunciation and fluency has followed the introduction of fully-automated standardized L2 speaking tests. Finally, the growing use of English as a Lingua Franca (ELF) in diverse international contexts brought about by globalization and technological advancements has catapulted the issue of defining an appropriate pronunciation standard to the frontline of assessment concerns (e.g., Davies 2013; Jenkins 2006), with discussions extending to pronunciation norms in lingua franca contexts for languages other than English (Kennedy et al. in press). New edited volumes (Isaacs & Trofimovich in press; Kang & Ginther forthcoming) are taking stock of these developments, fusing perspectives from research communities where there has, hitherto, been little communication.

This resurgence can be seen as part of a cycle, as there have been times in the past where pronunciation was at the forefront of language teaching, learning, and assessment (Isaacs 2014). The goal of this timeline is, therefore, to chart a clear historical trajectory of pronunciation assessment. In this, we will underscore how conceptualizations and practical implementations have evolved over time, with influences from teaching methodologies, theoretical frameworks, and seminal research that evidence (or in the case of newer pieces, have potential for) 'historical reverberation'. Throughout, we chart how new lines of inquiry may be instigating or reinforcing change in assessment practice, establishing links where possible between work in different eras.

The starting point for this endeavour requires defining the terms 'pronunciation' and 'assessment.' In the context of this review, 'pronunciation' is inclusive of both segmental (individual sounds) and suprasegmental (prosodic) features, although the assessment instruments cited (e.g., rating scales) have their own operational definitions that may diverge from this. Following Bachman (2004), the term 'assessment' refers to any systematic information gathering process used to foster an understanding of the phenomenon of interest (e.g., learners' ability or processes). Conversely, a 'test' denotes a particular type of assessment in which a performance is elicited and an inference/decision is made about that performance, usually on the basis of a test score. All tests are assessments, but not all assessments are tests—although tests are the most common type of formal assessment. Because tests tend to be higher-stakes and more ubiquitous than other assessment types, they are well-represented in the timeline, which includes both direct citations of assessment instruments, and the research and validation work which underpins their development and use. No timeline can be exhaustive, and English is overrepresented as the target language in the included entries.

Much of the focus of the timeline is on defining a suitable standard for assessing pronunciation (e.g., native like accuracy vs. intelligible/comprehensible speech), arriving at an adequate operational definition of pronunciation, or considering pronunciation in relation to some conception of aural-oral ability or communicative competence/communicative language ability. Although from a research perspective, the terms ‘intelligibility’ and ‘comprehensibility’ are frequently distinguished in how they are *operationalized* (e.g., using orthographic descriptions vs. rating scales in Derwing & Munro’s 2015 conception, although Smith & Nelson 1985, offer a different interpretation), these terms have not been used consistently in L2 speaking scales. The term used in the timeline is simply the one used by the author of the cited publication or assessment instrument.

Another prominent line of inquiry relates to reliability: how might pronunciation be objectively assessed? There is potential for individual differences in the characteristics of those scoring pronunciation assessments to unduly influence or bias the assessment, which raises issues of test fairness. Human raters can now be supplanted through the use of modern technology, which addresses the issue of human behavioural variability. However, machine scoring of speech is not without limitations, with automated scoring systems, as yet only able to robustly approximate human judgments on highly controlled L2 speaking tasks that yield predictable learner output (e.g., sentence read-aloud, construction, or repetition tasks). This has raised concerns within the assessment community about the narrowing of the L2 speaking construct using automated scoring (e.g., interactional patterns not captured; tasks relatively inauthentic; Chun 2006). Although improvements in technological capabilities offer much promise into the future, it is humans (not computers) who are relevant in the context of real-world communicative transactions. Relative to this standard, to which machine scoring will continue to be compared, there will always be limitations to what machines are able to measure and simulate (Isaacs 2016).

To capture the scope of topics and sources of influence, we organized papers into one or more of a range of themes. The themes were initially devised to cover four key areas: operational assessment systems, practitioner oriented guides, theoretical frameworks, and research studies/syntheses. However, given that peer-reviewed journal articles and other research publications constituted over two-thirds of the entries, the fourth area – research studies/syntheses – was split into three further categories: research investigating learner performance or development; research examining the role of non-linguistic factors in pronunciation assessment; and research which takes a broader view of assessment in relation to SLA or language pedagogy. The resulting themes are:

A: A language test or scoring system, including rating scales and automated assessments

B: A teaching methodology or assessment-oriented guide for language researchers and/or practitioners

C: A theoretical framework of language ability, knowledge, and/or processing

D: Research on defining or validating speech-related constructs, either as operationalized in an assessment instrument, or through investigations of human- or machine-derived linguistic measures in relation to learner performance or development

E: Research on the effects of nonlinguistic variables (e.g., attitudes, accent familiarity, age) on speakers' or listeners' test/task performance or on listeners' (raters'/examiners') judgments of speech

F: Lab or classroom-based L2 research incorporating a broader notion of assessment, including studies examining the effectiveness of pedagogical interventions

References

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Chun, C. W. (2006). An analysis of a language test for employment: The authenticity of the PhonePass test. *Language Assessment Quarterly* 3.3, 295–306.

Davies, A. (2013). *Native speakers and native users: Loss and gain*. Cambridge: Cambridge University Press.

Derwing, T. M. (2010). Utopian goals for pronunciation teaching. In J. Levis & K. LeVelle (eds.), *Proceedings of the 1st Pronunciation in Second Language Learning and Teaching Conference*. Ames, IA: Iowa State University, 24–37.

Derwing, T. M. & M. J. Munro (2015). *Pronunciation Fundamentals: Evidence-based perspectives for L2 teaching and research*. Amsterdam: John Benjamins.

Fulcher, G. (2015). Assessing second language speaking. *Language Teaching* 48.2, 198–216.

Harding, L. (2013). Pronunciation assessment. In C. A. Chapelle (ed.), *The encyclopedia of applied linguistics*. Hoboken, NJ: Wiley-Blackwell. Doi: 10.1002/9781405198431.wbeal0966.

Isaacs, T. (2014). Assessing pronunciation. In A. J. Kunnan (ed.), *The companion to language assessment* (vol. 1). Hoboken, NJ: Wiley-Blackwell, 140–155.

Isaacs, T. (2016). Assessing speaking. In D. Tsagari & J. Banerjee (eds.), *Handbook of second language assessment*. Berlin: DeGruyter Mouton, 131–146.

Jenkins, J. (2006). The spread of EIL: A testing time for testers. *ELT Journal*, 60.1, 42–50.

Kang, O. & A. Ginther (eds.) (forthcoming). *Assessment in second language pronunciation*. New York: Routledge.

Kennedy, S., J. Blanchet & D. Guénette (in press). Teacher-raters' assessments of French lingua franca pronunciation. In T. Isaacs & P. Trofimovich (eds.), *Second language* *pronunciation assessment: Interdisciplinary perspectives*. Bristol, UK: Multilingual Matters.

Levis, J. (2015). The Journal of Second Language Pronunciation: An essential step toward a disciplinary identity. *The Journal of Second Language Pronunciation* 1.1, 1–10.

Levis, J. & M. J. Munro (eds.) (2013). Phonetics and phonology [volume]. In C. A. Chapelle (ed.), *Encyclopedia of Applied Linguistics*. Hoboken, NJ: Wiley Blackwell.

Lim, G. S. & E. D. Galaczi (eds.) (forthcoming). Special Issue on Conceptualizing and operationalizing second language speaking assessment: Updating the construct for a new century. *Language Assessment Quarterly.*

Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.

Munro, M. J. & T. M. Derwing (2011). The foundations of accent and intelligibility in pronunciation research. *Language Teaching* 44.3, 316–327.

Purpura, J. E. (2016). Second and foreign language assessment. *The Modern Language Journal* 100.S1, 190–208.

Reed, M. & J. Levis (eds.) (2015). *The handbook of English pronunciation*. Malden, MA: Wiley-Blackwell.

Smith, L. E. & C. I. Nelson (1985). International intelligibility of English: Directions and resources. *World Englishes* 4.3, 333–342.

Year	References	Annotations	Theme
Circa 500 BCE	Judges, 12:5-6	This well-known passage from the Book of Judges describes a high-stakes pronunciation test, where fleeing Ephraimites were asked by the Gileadites at a border crossing to pronounce the word 'Shibboleth' in order to identify the Ephraimites, who were expected to pronounce the first syllable onset as /s/ instead of /ʃ/, with the Ephraimites' dialect lacking the /ʃ/ phoneme. On the basis of this test, individuals were either allowed to pass or were slaughtered. The shibboleth story has had far-reaching cultural ramifications, clearly showing that pronunciation assessment is not always a benign activity. Although typically less brutal, modern day shibboleth tests persist (McNamara & Roever 2006)[1].	**A**
1899	Sweet, H. (1899). *The practical study of languages: A guide for teachers and learners*. London: Dent.	In a rejection of the exclusive focus of the Grammar Translation method on the written medium, **Sweet** advocated 'basing all study of language on	**B**

		phonetics' (p. vii), placing phonetic transcription at the centre of teacher training, thereby reducing reliance on a native speaker to model correct pronunciation. In perhaps the earliest written reference to L2 intelligibility, Sweet argued for 'speaking with moderate fluency and sufficient accuracy of pronunciation to insure intelligibility' (p. 239). However, he also referred to mastery of the L2 sound system as a learning goal, long before evidence had emerged that native-like accuracy was elusive for most L2 learners (FLEGE 2005) and pedagogically incongruous with the goal of targeting intelligible speech (LEVIS 2005).	
1913	UCLES. (1913). *Certificate of Proficiency in English (CPE)*. Cambridge: UCLES.	SWEET's (1899) attempts to shift the instructional focus to speaking extended to formal testing in the development of the Certificate of Proficiency in English (CPE) for foreign language teachers, which included an oral paper and a written	**A**

		phonetics paper. Although the oral component is still integral to the Cambridge approach today, the Phonetics paper did not survive the first round of CPE revisions in 1932 (Weir et al. 2013)[2].	
1944	Kaulfers, W. V. (1944). Wartime development in modern-language achievement testing. *The Modern Language Journal* 28.2, 136–150.	In America, interest in assessing speaking was spurred by involvement in World War Two and the need to test communicative readiness for deployment in a foreign country. **Kaulfers**' article on wartime test development constituted perhaps the earliest attempt to operationalize intelligibility in a scale, with 'readily intelligible' as perceived by a 'literate native' listener at the highest level of the scale and 'unintelligible or no response' at the low end (p. 144). Most rating scales in use today similarly do not spell out which linguistic features specifically lead to breakdowns in understanding (ISAACS ET AL. 2015).	**A**

1958	Foreign Service Institute (1958). *FSI Proficiency Ratings*. Washington D.C.: Foreign Service Institute.	Oral assessment grew in importance during the Korean War, when it became clear that the US government needed a standard set of levels that could be used across languages to rate proficiency, spurring the development of the Foreign Service Institute (FSI) scales. It consisted of five scale criteria described over six levels, one of which was ‘accent.’ The top descriptor for the accent scale is ‘native pronunciation, with no trace of ‘foreign accent,’’ underscoring native-like accuracy rather than intelligibility at the highest level of achievement. The FSI scales ultimately led to the widespread use of the oral proficiency interview as a method for assessing speaking. They also directly influenced the development of the Interagency Language Roundtable (ILR) scales and the American Council on the Teaching of Foreign Languages (ACTFL) scales (see Chalhoub-Deville & Fulcher 2003)[3].	**A**

1960	Lambert, W.E., R. Hodgson, R. C. Gardner & S. Fillenbaum (1960). Evaluational reactions to spoken languages. *Journal of abnormal and social psychology* 60.1, 44–51.	As progress was made on L2 pronunciation assessment, a distinct line of research in social psychology led to the observation that attitudes toward speakers vary as a function of particular features of their pronunciation or speech style. This seminal study introduced the speaker evaluation paradigm through the 'matched-guise technique,' an experimental approach involving an actor mimicking native and/or L2 accents still widely in use today. Because listeners' social judgments about a speaker's personality or physical attributes are generally considered extraneous to the assessment of L2 speaking ability, it is important to minimize such attitudinal effects among pronunciation assessors. At the same time, this study highlighted that pronunciation assessment (e.g., judgements of competence based on speech patterns)	**E**

		may occur daily across many social situations.	
1961	Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. London: Longman.	In **Lado**'s seminal book on practicalities in designing, administering, and scoring language tests, pronunciation is the most comprehensively covered language component, with chapters on testing the perception and production of segments, stress, and intonation. One challenge he articulated was the 'insoluble' problem of using intelligibility as the pronunciation assessment standard, including the issue of 'what natives are to be used as touchstones' (p. 79) in judging whether or not speech is intelligible. Subsequent research on rater effects has revealed the importance of this consideration (e.g., CAREY, MANNELL & DUNN 2011).	**B**
1980	Canale, M. & M. Swain (1980). Theoretical bases of communicative approaches to second language teaching and	Although the communicative turn in language teaching and testing had begun in the late 1960s, **Canale & Swain**'s model of communicative	**C**

	testing. *Applied Linguistics* 1.1, 1–57.	competence, which consists of grammatical, sociolinguistic, and strategic competence, provided the theoretical rigor upon which subsequent work could be built (e.g., BACHMAN 1990). Pronunciation falls under grammatical competence, where it is referred to as knowledge of phonological rules. While there is scope within this approach to explore the role of, for example, intonation in making sociolinguistically appropriate utterances, it seems fair to say that the importance of pronunciation in the model is minimal, signalling a shift away from pronunciation throughout the 1980s and early 1990s, as buttressed by KRASHEN'S (1982) views about formal instruction being ineffective or a hindrance.	
1982	Krashen, S. (1982). *Principles and practice in second language acquisition*. Oxford: Pergamon Press.	Although acknowledging the dearth of research on instructional effects, **Krashen** argued that explicit pronunciation teaching (e.g., pattern-drills, repetitive activities) either did	**B, C**

		not improve learners’ pronunciation ability, or was inferior to communicatively-oriented instruction. The implication was that learners can acquire pronunciation by osmosis, a view which contributed to its marginalization in classroom teaching and research and its side-lining in assessment circles for decades (ISAACS & TROFIMOVICH 2012).	
1987	Fayer, J. M. & E. Krasinski (1987). Native and nonnative judgments of intelligibility and irritation. *Language Learning* 37.3, 313–326.	One of the key variables in pronunciation assessment is the assessor. Someone needs to judge the correctness or appropriateness of pronunciation, and that person comes with individual biases. **Fayer & Krasinski** presented one of the earliest studies of rater bias in their investigation of native and non-native listeners’ judgements of intelligibility, finding that non-native listeners found their own accent more annoying than did native listeners. This study paved the way for future research on rater	**E**

		effects in formal and informal pronunciation assessments.	
1989	Buck, G. (1989). Written tests of pronunciation: Do they work? *ELT Journal* 43.1, 50–56.	As a less resource-intensive alternative to administering and scoring oral pronunciation tests, LADO (1961) proposed using paper-and-pencil pronunciation items, hypothesizing that written scores would strongly correlate with test-takers' oral pronunciation. **Buck** tested this hypothesis using a test modelled on LADO's written item prototypes and found unacceptably low correlations between the written test scores and ratings of test-takers' oral pronunciation. He also reported 'catastrophically low reliabilities' among the items (p. 54), concluding that the test was an invalid and unreliable measure of pronunciation production. Despite these concerns, written items modelled on LADO's (1961) blueprints are still in use in the high-stakes English language National	**D**

		Center Test for University Admissions in Japan (Isaacs 2014).	
1989	Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.	There is, as yet, no comprehensive or falsifiable theoretical model of pronunciation assessment. **Levelt**'s speech production model, which posits the processing components and knowledge sources involved in conceptualizing, formulating, and articulating speech from a first language (L1) cognitive perspective, has been featured in work on L2 speech perception, production, and the design of standardized speaking tests. However, its integration into SLA-oriented L2 pronunciation research and applications for psycholinguistically-oriented pronunciation assessment have yet to be fully realized.	**C**
1990	Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.	Building on Canale & Swain (1980), **Bachman**'s communicative language ability framework has arguably been the dominant theoretical view for conceptualizing L2 ability in the	**B, C**

		language assessment field since its publication. However, his coupling of ‘phonology/graphology,’ where the latter term refers to the legibility of handwriting, is unexplained and underconceptualized—likely a remnant from Lado’s (1961) skills-and-components model.	
1992	Anderson-Hsieh, J., R. Johnson & K. Koehler (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning* 42.4, 529–555.	This empirical study revealed that prosodic errors have a stronger effect on intelligible pronunciation than do segmental or syllable structure errors. The study led the way for further research on the relationship between ratings of different pronunciation dimensions and the quantifiable features of those dimensions in speech samples (e.g., Kang 2010).	**D**
1992	Rubin, D. L. (1992). Nonlanguage factors affecting undergraduates' judgments of non-native English-speaking teaching assistants. *Research in*	Building on earlier sociolinguistic studies mostly examining attitudes toward different L1 regional accents (e.g., Lambert et al. 1960), **Rubin** demonstrated that listeners’ perceptions of L2 speech are mediated by their preconceptions of talkers. In	**E**

	Higher Education 33.4, 511–531.	his study, American undergraduate students who listened to a recording of a native English speaker while viewing the photo of an Asian instructor, understood less of the lecture than did a comparison group who listened to the same recording while viewing the photo of a Caucasian instructor. This study was a harbinger of further L2 pronunciation research on construct-irrelevant sources of variance (i.e., variables extraneous to the speech productions being measured) and their potential to bias listeners' assessments (Kang & Rubin 2009)[4].	

1995	Munro, M. J. & T. M. Derwing (1995). Foreign accent, intelligibility and comprehensibility in the speech of second language learners. *Language Learning* 45.1, 73−97.	**Munro & Derwing**'s pioneering study, which opened-up a rich line of enquiry, introduced conceptually clear operational definitions of the terms 'intelligibility,' 'comprehensibility,' and 'accentedness,' which have been widely (although not universally) used in L2 pronunciation research (Isaacs & Thomson 2012). They also demonstrated that the constructs of intelligibility and comprehensibility cannot be equated with accentedness. Historically, several rating scales have conflated these partially independent dimensions (e.g., FSI) and this is still the case in scales in use today (e.g., CEFR Phonological control scale).	**D**

1995	Flege, J. E., M. J. Munro & I. R. A. Mackay (1995). Factors affecting strength of perceived foreign accent in a second language. *Journal of the Acoustical Society of America* 97.5, 3125−3134.	In one of the largest age-related studies, **Flege et al.** found a strong monotonic relationship between age of arrival in the target language country, which was used as an index of age of L2 learning, and perceived L2 accent, with earlier learners receiving less accented or more native-like ratings than speakers who had learned the L2 later in life. Some listeners were able to detect an L2 accent in speakers well before what is traditionally considered to be the critical period (< 4 years), providing indirect evidence for the sensitivity of untrained raters in distinguishing native- from non-native speech. An implication is that acquiring native-like accuracy is an unrealistic goal for pronunciation instruction and, by implication, assessment.	**E**
1996	Celce-Murcia, M., D. Brinton & J. Goodwin (1996). *Teaching pronunciation: A reference for teachers of*	Among the most well-known and comprehensive pronunciation texts for classroom teachers, **Celce-Murcia, Brinton & Goodwin** provide in-depth	**B**

	English to speakers of other languages. Cambridge: Cambridge University Press.	coverage of pronunciation assessment in the final chapter of their book. Particularly impressive is the focus on diagnostic approaches to pronunciation assessment well before the current diagnostic assessment zeitgeist.	
1999	Bernstein, J. (1999). *PhonePass testing: Structure and construct*. Menlo Park, CA: Ordinate Corporation.	The emergence of **PhonePass** in the 1990s signified the first steps for the language assessment field into the world of automated scoring of L2 speech. This was achieved using an automatic speech recognition (ASR) system, initially trained on a large sample of speech ratings conducted by human listeners, to develop the scoring algorithm. Pronunciation (particularly segmentals) and fluency are key parts of the construct, as the ASR system is heavily dependent on spectral and durational measures produced on a range of controlled L2 speech tasks. PhonePass demonstrated high correlations with scores from more traditional language proficiency	**A**

		instruments, suggesting that speaking assessment might be possible through cheap and efficient methods that are readily available to stakeholders (e.g., PhonePass was administered over the phone). The PhonePass technology, originally developed by Ordinate, was acquired by Pearson in 2008, and the patented system is now used across the Versant suite of language tests and other Pearson products (e.g., Pearson Test of English Academic; Bernstein et al. 2010)[5].	
2000	Cucchiarini, C., H. Strik & L. Boves (2000). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *Journal of the Acoustical Society of America* 107.2, 989–999.	**Cucchiarini et al.**'s experiment using read-aloud productions of L2 learners of Dutch provides evidence that temporal measures (e.g., articulation rate), derived using an automatic speech recognizer, are reliable and sufficiently strongly correlated with 'expert' human ratings (assessed by phoneticians/speech therapists) to be useful for developing automated assessments of L2 speech. This is a rare study in its discussion of	**A, D**

		assessment and is part of a larger body of work examining the efficacy of using machine-generated pronunciation feedback in computer-assisted language learning.	
2000	Jenkins, J. (2000). *The phonology of English as an international language.* Oxford: Oxford University Press.	**Jenkins'** (2000) book represented something of a revolution in pronunciation learning and teaching, shifting the focus toward intelligibility in English as a Lingua Franca (ELF) settings—that is, contexts where language users who do not share an L1 use English as the common language of communication. Jenkins developed a set of pronunciation features called the lingua franca core (LFC) which she viewed as crucial for intelligibility in ELF contexts, excluding features which were considered unimportant for intelligibility (e.g., connected speech). While the LFC has been critiqued for numerous reasons, including having been derived from a limited dataset (Isaacs 2014), there is no doubting its	**B, D**

		influence as the genesis for a program of research and critical pedagogy. In assessment, the ideas have yet to be implemented by large exam boards but become relevant when considering pronunciation in paired/group oral assessments, where Jenkins' work on accommodation (i.e., convergence/divergence of interlocutors' pronunciation patterns during interactions) could be a consideration, for example, in same- versus different-L1 pairings.	
2001	Council of Europe (2001). *Common European Framework of Reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.	The **Council of Europe**'s Common European Framework of Reference (CEFR), which describes language ability across six reference levels, excludes pronunciation from its global descriptors, which implies that pronunciation is unimportant for measuring language proficiency, making it a stealth factor in scoring (Isaacs 2014). The CEFR Phonological control scale, one of six additional fine-grained scales	**A, B, C**

		targeting ‘linguistic competences,’ conflates the constructs of strength of L2 accent and ease of understanding, despite the lack of empirical basis for this (Munro & Derwing 1995). At the time that this research timeline went to print, efforts to revise the Phonological control descriptors were underway.	
2003	Bent, T. & A. R. Bradlow (2003). The interlanguage speech intelligibility benefit. *Journal of the Acoustical Society of America* 114.3, 1600–1610.	**Bent & Bradlow’s** study demonstrated that listeners might receive an intelligibility advantage if they share a speaker’s L1, spawning a growing body of subsequent research on the topic (e.g., Harding 2012). Their finding raises the prospect of rater bias if an assessor shares (or is highly familiar with) a speaker’s accent—a variable which might need to be controlled for or screened in rater selection for high-stakes tests and research studies alike (Winke, Gass & Myford 2013)[6]. It also problematizes the use of speakers with different accents in L2 listening tests	**E**

		intended for test-takers from mixed L1 backgrounds, since listeners' familiarity with the accent used in the prompt could lead to greater item difficulty (Ockey & French 2014)[7].	
2005	Educational Testing Service (ETS). (2005). *Test of English as a Foreign Language internet-based test (iBT)*. Princeton, NJ: ETS.	The original paper-based TOEFL test was first introduced in 1964. However, it was not until its launch as the TOEFL internet-based test (iBT) in 2005—after two major revisions—that a mandatory speaking section was included. Prior to this, proof of proficiency for university admissions screening and, in some cases, employment as an international teaching assistant had no speaking requirement (ISAACS, 2008). In the TOEFL iBT analytic scoring rubric, pronunciation (e.g., intelligibility, stress, intonation) and fluency features are assessed under the 'delivery' criterion. Given the global reach of the TOEFL, the introduction of pronunciation as a measured ability is	**A**

		likely to have had a major washback effect in classrooms around the world.	
2005	Levis, J. (ed.) (2005). Special issue on pronunciation. *TESOL Quarterly* 39.3.	The publication of *TESOL Quarterly's* groundbreaking special issue on pronunciation featured contributions on the incompatibility of targeting accent reduction versus intelligibility in pronunciation instruction (which Levis described as stemming from the 'nativeness principle' versus 'intelligibility principle,' respectively, in his article), perspectives on JENKINS' (2000) LFC, the effects of selected pronunciation features on intelligibility, and listeners' social evaluations of L2 accents. Although there were no articles directly focused on pronunciation assessment, the reintegration of pronunciation into mainstream English language research and teaching, as attested by this special issue in a wide-circulation journal, led the way for the uptake of such issues in assessment-related	**D, E, F**

		research (e.g., ISAACS & TROFIMOVICH, 2012; KANG 2012).	
2008	Isaacs, T. (2008). Towards defining a valid assessment criterion of pronunciation proficiency in non-native English speaking graduate students. *Canadian Modern Language Review* 64.4, 555–580.	**Isaacs**' (2008) research was among the first of the assessment-focused pronunciation studies to be published in the wake of LEVIS (2005), and was unique in its melding together of more recent conceptualizations of intelligibility with the key question of language test design: validity. Specifically, she investigated whether intelligibility was a sufficiently broad pronunciation construct for screening international teaching assistants, and found that, in this case, it was not.	**D**
2010	Kang, O. (2010). Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System* 38.2, 301–315.	**Kang**'s article on the relative contribution of acoustic and temporal measures on native listeners' comprehensibility and accentedness judgments is among the first of a collection of assessment-oriented studies to use Praat, a freely-available speech analysis application widely used by phoneticians and applied linguists. Subsequent publications	**D**

		written primarily for a language assessment audience addressed the implications of using such objectively-derived measures for automated scoring (e.g., Kang & Pickering 2014)[8].	
2010	Xi, X. (2010). Special issue on automated scoring and feedback systems for language assessment and learning. *Language Testing* 27.3.	Following the acquisition of the PhonePass technology by Pearson and in the wake of the rollout of their fully-automated tests, there had been increasing interest in ASR within assessment circles. This special issue of *Language Testing* was pioneering in drawing together specialists in automated scoring, with several articles reporting on speech recognition innovations, with applications for pronunciation assessment and feedback provision to test-takers.	**D**
2011	Carey, M. D., R. H. Mannell & P. K. Dunn (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral	Situated in a growing volume of research investigating rater familiarity effects on L2 speaking assessments, **Carey et al.** examined effects on pronunciation scoring specifically,	**E**

	proficiency interviews? *Language Testing* 28.2, 201–219.	showing that familiarity may have a noticeable effect on pronunciation ratings even among trained IELTS examiners.	
2012	Harding, L. (2012). Accent, listening assessment and the potential for a shared-L1 advantage: A DIF perspective. *Language Testing* 29.2, 163–180.	Bringing the issues of pronunciation and listening assessment together, **Harding** extended BENT & BRADLOW's (2003) 'interlanguage speech intelligibility benefit' to L2 listening tests, demonstrating some evidence of L1-mediated listener bias using differential item functioning. This article argues for the need to expose test-takers to different varieties of English in listening assessments, and that research attention should turn to developing suitable methods for selecting diverse-accented speakers with equivalent intelligibility for listening input.	**E**
2012	Isaacs, T. & P. Trofimovich (2012). 'Deconstructing' comprehensibility: Identifying the linguistic influences on listeners' L2	Building on previous research by MUNRO & DERWING (1995) and KANG (2010) on examining correlations between linguistic measures and L2 comprehensibility ratings, **Isaacs &**	**D**

	comprehensibility ratings. *Studies in Second Language Acquisition* 34.3, 475–505.	**Trofimovich**’s work was the first of a series of studies to show that comprehensibility is related to a wide range of linguistic domains, including segmental, prosodic, temporal, lexicogrammatical, and discourse-level measures. They also demonstrated the potential for operationalizing comprehensibility in an empirically-based rating scale to offset the limitations of intuitively-developed scales, opening up the potential for further work on examining the generalizability of comprehensibility scale criteria across test-takers’ L1 background and task type (e.g., Crowther et al. 2015)[9].	
2012	Saito, K. & R. Lyster (2012). Effects of form-focused instruction and corrective feedback on L2 pronunciation development of /ɹ/ by Japanese learners of English. *Language Learning* 62.2, 595–633.	**Saito & Lyster**’s article was the first to investigate corrective feedback effects in relation to pronunciation learning in SLA research. The major finding was that form-focused instruction needed to be accompanied by systematic, incidental correction of pronunciation errors (recasts) to be	**F**

		effective. This study is relevant to the growing body of classroom-based L2 assessment research that views assessment (including feedback) as integral to teaching and learning. It also contributes to the relatively small body of research on the effects of instructional treatments on ‘fossilized’ error types that could interfere with intelligibility (Saito 2012)[10].	
2013	Isaacs, T. & R. I. Thomson (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly* 10.2, 135–159.	Since Munro & Derwing (1995), judgements of pronunciation in SLA research have typically been measured on Likert-type comprehensibility, accentedness, and/or fluency scales. While these scales have become ubiquitous, they have rarely been scrutinized from a psychometric perspective. **Isaacs & Thomson** examine optimal scale length and also the variable of rater experience. The results problematize the use of these scales in SLA research, demonstrating that a language assessment perspective	**D**

		on research methodology can be fruitful.	
2014	Lee, J., J. Jang & L. Plonsky (2014). The effectiveness of second language pronunciation instruction: A meta-analysis. *Applied Linguistics* 36.3, 345–366. doi:10.1093/applin/amu040	To counter decades of discourse on the neglect of pronunciation in L2 research and pedagogy, reviews and meta-analyses centring on the instructional efficacy and targets of L2 pronunciation instruction in SLA research began to appear in the second decade of the 21st century, enabling a critique of methodology, including for assessment-relevant variables (e.g., task type, mode of delivery, feedback provision). For example, **Lee et al.**'s evidence synthesis revealed medium to large positive effect sizes for pronunciation instruction, with stronger effects in lab than classroom-based studies. This finding provides counterevidence to KRASHEN's (1982) claim that formal instruction on linguistic forms is counterproductive.	**F**
2015	Isaacs, T., P. Trofimovich, G. Yu & B. M. Chereau (2015). Examining the linguistic	In a study on the revised IELTS Pronunciation scale, following its expansion from a four- to nine-point	**D**

	aspects of speech that most efficiently discriminate between upper levels of the revised IELTS pronunciation scale. *IELTS research reports online series*, 4.	scale in 2008, **Isaacs et al.** found that identifying a single linguistic measure that distinguishes between adjacent IELTS Pronunciation levels is elusive. However, they made several practical recommendations based on accredited examiners' ratings and perspectives, including reordering descriptors within bands from more global (comprehensibility) to more discrete features, delineating pronunciation criteria at Bands 5 and 7 to implement a clearer division and lessen examiners' cognitive load, and minimizing background noise at test centres if comprehensibility is among the assessed criteria, as this is a potential confound. The study confirmed previous findings that examiners perceive Pronunciation as the most difficult IELTS Speaking subscale to rate (Yates, Zielinski & Pryor 2011)[11], making the need for generating more precise descriptors all the more pressing.	

2016	Trofimovich, P., T. Isaacs, S. Kennedy, K. Saito, & D. Crowther (2016). Flawed self-assessment: Investigating self- and other-perception of second language speech. *Bilingualism: Language and Cognition* 19.1, 122–140.	One area that is underrepresented in this timeline relates to work on peer- and self-assessment of L2 pronunciation. **Trofimovich et al.**’s study partially addresses this gap, examining L2 learners’ self-assessments of accentedness and comprehensibility in relation to linguistic measures rated by native speakers. The major finding was that L2 learners who are at the low end of the accentedness and comprehensibility continuum tended to overestimate their performance whereas high ability learners tended to underestimated it. The discrepancies between self- and other-assessment were linked to segmental and prosodic measures rather than to lexical, grammatical, or discourse-level measures. The study opens up the potential for further exploration, including pairing teacher- or peer-assessments or more objective pronunciation measures with self-	**D, F**

		assessments to heighten leaners' awareness and help them develop less distorted views of their own abilities.	
2017	Isaacs, T. & P. Trofimovich (eds.) (in press). *Second language pronunciation assessment: Interdisciplinary perspectives*. Bristol, UK: Multilingual Matters.	The central contribution of this first edited collection on pronunciation assessment is bringing together perspectives from different research communities with little crossover (assessment, psycholinguistics, sociolinguistics, lingua franca, SLA, and speech sciences) to develop a baseline understanding of principles, terminology, and priorities for future pronunciation assessment research, including drawing on insights from assessing other skills (e.g., writing, listening). Content coverage of the book is non-exhaustive and a notable omission is a chapter on automated assessment (BERNSTEIN 1999; XI, 2010)—a gap that a forthcoming edited collection on pronunciation assessment by Kang & Ginther is likely to fill.	**D, E, F**

[1] McNamara, T. F. & C. Roever (2006). *Language testing: The social dimension*. Malden,

MA: Blackwell.

[2] Weir, C. J., I. Vidaković & E. Galaczi (2013). *Measured constructs: A history of Cambridge English language examinations 1913–2012*. Cambridge: Cambridge University Press.

[3] Chalhoub-Deville, M. & G. Fulcher (2003). The oral proficiency interview: A research agenda. *Foreign Language Annals* 36.4, 498–506.

[4] Kang, O. & D. L. Rubin (2009). Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. *Journal of Language and Social Psychology*, 28.4, 441–456.

[5] Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*, 27.3, 355–377.

[6] Winke, P., S. Gass & C. Myford (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing* 30.2, 231–252.

[7] Ockey, G. & R. French (2014). From one to multiple accents on a test of L2 listening comprehension. *Applied Linguistics*. Advance access doi: 10.1093/applin/amu060

[8] Kang, O. & L. Pickering (2014). Using acoustic and temporal analysis for assessing speaking. In A. J. Kunnan (ed.), *The companion to Language Assessment* (vol. 2). Hoboken, NJ: Wiley-Blackwell, 1047–1062.

[9] Crowther, D., Trofimovich, P., Isaacs, T. & Saito, K. (2015). Does speaking task affect second language comprehensibility? *Modern Language Journal* 99.1, 80–95.

[10] Saito, K. (2012). Effects of instruction on L2 pronunciation development: A synthesis of 15 quasi-experimental intervention studies. *TESOL Quarterly* 46.4, 842–854.

[11] Yates, L., E. Zielinski & E. Pryor (2011). *The assessment of pronunciation and the new IELTS Pronunciation Scale*. In J. Osborne (ed.), IELTS Research Reports (vol. 12). Melbourne: IDP IELTS Australia, 23–68.