8

Assessing Pronunciation

Talia Isaacs
University of Bristol, England

Introduction

Accents are one of the most perceptually salient aspects of spoken language. Previous research has shown that linguistically untrained listeners are able to distinguish between native and non-native speakers under nonoptimal experimental conditions, including when the speech is played backwards (Munro, Derwing, & Burgess, 2010) or when it is in a language that listeners do not understand (Major, 2007). In fact, one of the earliest documented examples of language testing, the biblical Shibboleth test described in the Book of Judges, involved testing the identity of members of warring tribes based on whether they pronounced the word *shibboleth* 'sheave of wheat' with a /ʃ/ or a /s/ sound at syllable onset, with fatal consequences if the "wrong" pronunciation betrayed their enemy status (Spolsky, 1995). In modern times, a less brutal but still high stakes example is the use of so-called experts' analyses of speech to determine the legitimacy of asylum seekers' claims based on their *perceived* group identity (Fraser, 2009). Of course, such identity tests are far from foolproof, can lead to erroneous conclusions that could inform high stakes decisions, and raise concerns about fairness. It is often unclear, for example, whether it is aspects of the speech signal that trigger unfavorable listener responses, or whether listener expectations that arise as a result of linguistic stereotyping lead listeners to assign qualities to the speech that are absent or distorted (Kang & Rubin, 2009).

Foreign accents tend to receive a disproportionate amount of attention precisely due to their perceptual salience. Despite the enduring reference to the native speaker as the "gold standard" of language knowledge (Levis, 2005), eradicating traces of a foreign accent is widely viewed by applied linguists as an unsuitable goal for L2 pronunciation instruction for several reasons. First, native-like attainment of phonology is an unrealistic goal for most adult L2 learners, not least

The Companion to Language Assessment, First Edition. Edited by Antony John Kunnan.

DOI: 10.1002/9781118411360.wbcla012

possibly an undesirable goal for L2 speakers, since accent and identity are intertwined (Gatbonton & Trofimovich, 2008). Second, L2 speakers do not need to sound like native speakers to fully integrate into society or successfully carry out their academic or professional tasks (Derwing & Munro, 2009). Third, the global spread of English and its emergence as the international lingua franca renders conformity to native speaker norms inappropriate in many EFL settings (Jenkins, 2002). In fact, many native English speakers themselves do not speak prestige (standard) varieties of English (e.g., Received Pronunciation, General American English). For all of these reasons, having a native-like accent is an unsuitable benchmark for pronunciation assessment in the vast majority of language use contexts.

The emerging consensus among applied linguists is that what really counts in oral communication is not accent reduction or attaining a native-like standard but rather simply being understandable to one's interlocutors and able to get the message across (Jenkins, 2002). In fact, over a decade of L2 pronunciation research has shown that having an L2 accent does not *necessarily* preclude L2 speech from being perfectly understandable, although it might. It is in cases when the presence of an L2 accent does impede listener understanding that explicit instruction is most needed to address learners' pronunciation difficulties (Derwing & Munro, 2009).

The theme of defining and operationalizing an appropriate assessment criterion for L2 pronunciation permeates this chapter. After providing reasons for the exclusion of pronunciation from L2 classrooms and its marginalization from mainstream L2 assessment research over the past several decades, the role of pronunciation in theoretical models of communicative competence and in L2 oral proficiency scales will be examined. Next, existing empirical evidence on the pronunciation features that should be taught and, by implication, tested will be considered, and research on individual differences in rater characteristics that could influence their judgments of L2 pronunciation will be discussed. The chapter will conclude with future directions in L2 pronunciation assessment research, with particular emphasis on technological innovations.

Previous Views or Conceptualization

In 1957, the English linguist J. R. Firth famously wrote, "you shall know a word by the company it keeps" (p. 11). A quick perusal of the past several decades of L2 pronunciation research reveals that "pronunciation" has kept close company with the term "neglect" (e.g., Derwing & Munro, 2009). This disparaging association generally refers to the devaluation of pronunciation by some communicative proponents and its resulting de-emphasis in ESL classrooms. One reason for the exclusion of pronunciation from L2 communicative teaching is the belief that an overt focus on pronunciation is extraneous to helping learners achieve communicative competence (Celce-Murcia, Brinton, Goodwin, & Griner, 2010). To counter this view, Morley (1991) argued that "intelligible pronunciation is an essential component of communicative competence" and that "ignoring students' pronunciation needs is an abrogation of professional responsibility" (pp. 488–9), since

poor pronunciation can be professionally and socially disadvantageous to L2 speakers. There is also evidence that adult L2 learners with "fossilized" pronunciation benefit from explicit pronunciation instruction (Derwing & Munro, 2009) and that a focus on pronunciation can be embedded in genuinely communicative activities (Trofimovich & Gatbonton, 2006).

Although the subject of L2 pronunciation *teaching* conjures up reference to neglect, there is at least a body of literature documenting this neglect. Not the same can be said about L2 pronunciation *assessment*, which, with the exception of literature on automated scoring, has been essentially dropped from the research agenda since the publication of Lado's seminal book, *Language Testing*, over half a century ago (1961). In what remains the most comprehensive treatment of L2 pronunciation assessment to date, Lado devoted separate chapters to testing L2 learners' perception and production of individual sounds, stress, and intonation, offering concrete guidelines on item construction and test administration. Some of Lado's views on L2 pronunciation are timely, including challenges in defining a standard of intelligible (i.e., easily understandable) pronunciation. However, other ideas are clearly outdated. For example, operating under the premise that "language is a system of habits of communication" (p. 22), Lado held that where differences exist between sounds in the learner's first language (L1) and the target language, there will be problems, and these need to be systematically tested. However, predicting learner difficulties appears to be more nuanced than a simple inventory of differences between the L1 and L2 can account for. There is growing evidence, for example, that the accurate perception and production of L2 segments (i.e., vowel or consonant sounds) is mediated by learners' *perceptions* of how different a given sound is from their existing L1 sound categories (Flege, Schirru, & MacKay, 2003). In general, accurate perception/production is more likely if the learner does not perceptually identify an L2 sound with any L1 sounds. This is because, if no difference is perceived, the learner will simply substitute the L1 sound for the L2 sound. In addition, contextual factors such as phonetic environment and lexical frequency also contribute to learner performance (Flege et al., 2003). Clearly, Lado's (1961) view that differences between L1 and L2 phoneme inventories should form the basis of L2 pronunciation tests oversimplifies the situation.

Due to advancements in language testing and speech sciences research, there is an urgent need for an updated guide on L2 pronunciation assessment and item writing. As reported above, Lado's work is the only extensive treatment on the subject. Therefore, several decades later, this reference remains the starting point for any discussion on L2 pronunciation assessment and, thus, features prominently in this chapter.

Lado expressed concern about the subjective scoring of test takers' speech and proposed the use of more objective paper and pencil tests as an alternative to assessing test takers' L2 pronunciation production (e.g., using multiple choice). Such written tests have the advantage of facilitating the testing of large numbers of students without the added time or expense of recording and storing speech samples or double marking them. The National Centre Test in Japan, a gatekeeping test for university admissions, uses decontextualized written items of the sort that Lado proposed to test oral pronunciation skills (see http://school.js88.com/

sd_article/dai/dai_center_data/pdf/2010Eng.pdf). The pronunciation component of the 2010 version consists of (a) segmental items, in which the test taker selects the word where the underlined sound is pronounced differently from the others (e.g., b*oo*t, g*oo*se, pr*oo*f, w*oo*l; the vowel sound in 'wool' /ʊ/ is different from the /u/ sound in the other choices); and (b) word stress items, in which the test taker selects the word that follows the same primary stress pattern as the item in the prompt (e.g., fortunately → appreciate, elevator, manufacture, sympathetic; both 'fortunately' and 'elevator' have primary stress on the first syllable).

In an empirical study on retired National Centre Test items entitled "Written Tests of Pronunciation: Do They Work?" conducted in a Japanese junior college, Buck (1989) found no evidence that they do. First, internal consistency coefficients (KR-20) for six pronunciation subtests were unacceptably low (range: –.89 to .54) as were correlations between scores on the written items and on test takers' oral productions of those items (.25 to .50). Correlations with read-aloud and extemporaneous speech task ratings were even lower (.18 to .43). Several decades after the publication of Lado's (1961) book and Buck's (1989) article, there is still no empirical evidence that written pronunciation items constitute a reliable or valid measure of L2 pronunciation speaking ability. In the absence of such evidence, the use of paper and pencil tests for oral production should be discontinued, particularly when they are being used for high stakes purposes.

Current Views or Conceptualization

Theoretical Conceptualization

The field of language testing has moved beyond Lado's (1961) focus on discrete-point testing and theoretical view of language as consisting of separate skills (speaking, reading, writing, listening) and components (e.g., vocabulary, grammar, pronunciation) toward expanded notions of communicative competence and communicative language ability. However, the assessment of L2 pronunciation has been left behind, with communicatively oriented theoretical frameworks not adequately accounting for the role of pronunciation. In Bachman's (1990) influential communicative language ability framework, for example "phonology/graphology" appears to be a carryover from the skills-and-components models of the early 1960s (Lado, 1961). However, the logic of pairing "phonology" with "graphology" (legibility of handwriting) is unclear. Notably, Bachman and Palmer's (1982) multitrait-multimethod study, which informed the development of Bachman's (1990) model, omitted the "phonology/graphology" variable from the analysis even though it was hypothesized to be an integral part of grammatical competence. This is because the authors claimed that phonology/graphology functions more as a channel than as a component, since pronunciation accuracy (and legibility) cannot be examined below a critical level at which communication breaks down. Bachman's reincorporation of phonology/graphology as a component in his 1990 model without explanation demonstrates the need for greater clarity on the role of pronunciation in communicative models.

In the L2 pronunciation literature, Levis has characterized two "competing ideologies" or "contradictory principles" that have long governed research and pedagogical practice (2005, p. 370). The first principle, the "nativeness principle," holds that the aim of pronunciation instruction should be to help L2 learners achieve native-like pronunciation by reducing L1 traces from their speech. The construct of "accentedness" in the L2 pronunciation literature, defined as listeners' *perceptions* of how different an L2 utterance sounds from the native-speaker norm (measured using rating scales), aligns with this principle. The second principle, the "intelligibility principle," holds that the goal of L2 pronunciation instruction should simply be to help L2 learners be understandable to their interlocutors—a view that most L2 researchers endorse and which is also "key to pronunciation assessment" (Levis, 2006, p. 252). However, the issue that Lado (1961) raised of "intelligible to whom" still resonates. To complicate matters, some scholars have depicted intelligibility as interactional between the speaker and the listener, whereas others have underscored that intelligibility is principally "hearer-based," or a property of the listener (Fayer & Krasinski, 1987, p. 313). Still others have criticized the burden that is implicitly placed on L2 speakers to achieve intelligibility, arguing that native speakers need to assume their share of the communicative responsibility (Lindemann, 2002).

Part of the problem is that intelligibility has been defined and measured in multifarious ways, which makes cross-study comparisons difficult (Isaacs, 2008). At least some of the confusion lies in the existence of broad and narrow definitions of the term. In its broad meaning, "intelligibility" refers to listeners' ability to understand L2 speech and is synonymous with "comprehensibility" (Levis, 2006). Reference to intelligibility as the appropriate goal of L2 pronunciation instruction and assessment conforms to this broad meaning. In its narrower sense, Derwing and Munro's (1997) conceptually clear definitional distinction between intelligibility and comprehensibility, which is increasingly pervasive in L2 pronunciation research, is useful to examine. Derwing and Munro define intelligibility as the amount of speech that listeners are able to understand (i.e., listeners' *actual* understanding). This construct is most often operationalized by computing the proportion of an L2 learner's utterance that the listener correctly orthographically transcribes. In contrast, comprehensibility, the more subjective measure, is defined as listeners' *perceptions* of how easily they understand L2 speech. This construct is operationalized by having raters record the degree to which they can understand L2 speech on a rating scale. Thus, comprehensibility, in its narrow definition, is instrumentally defined in that it necessitates a scale (i.e., a measurement apparatus) in the same way that measuring temperature necessitates a thermometer. That is, what distinguishes narrowly defined intelligibility from comprehensibility is not theory but, rather, the way these constructs have been operationalized. Hereafter, the term "comprehensibility" will therefore be used in its narrow sense whenever the notion of understandability is evoked in rating scales, with the exception of when the original wording from a given rating descriptor is retained. The term "intelligibility" will be used in both its broad and its narrow senses in the remainder of this chapter and the sense in which it is being used will be specified. The role of pronunciation in general and comprehensibility and accentedness in particular in current L2 speaking scales is the subject of the next section.

The Role of Pronunciation in Current Rating Scales

Theory often informs rating scale development. Because the theoretical basis for L2 pronunciation in communicative frameworks is weak as is our understanding of major holistic constructs, it follows that there are numerous shortcomings in the way pronunciation has been modeled in existing rating scales. First, pronunciation is sometimes omitted as a rating criterion. For example, pronunciation was excluded from the Common European Framework of Reference benchmark levels due to the high misfit values (i.e., substantial unmodeled variance) obtained for the pronunciation descriptors (North, 2000). Other scales that do include pronunciation only incorporate this criterion haphazardly. For instance, in the 10-level ACTFL oral Proficiency Guidelines (1 = novice low, 10 = superior), pronunciation is referred to in levels 1, 3, 4, and 5 of the scale but is entirely omitted from level 2 (novice mid). It is unlikely that pronunciation does not contribute to L2 oral proficiency at this precise point of the scale (level 2) when it is relevant at both neighboring levels. The inconsistency of reference to pronunciation or its exclusion altogether implies that pronunciation is not an important component of L2 speaking proficiency, making it likely that "pronunciation will become a stealth factor in ratings and a source of unsystematic variation in the test" (Levis, 2006, p. 245).

Another limitation of current scales is that their descriptors are often too vague to articulate a coherent construct. For example, in the public version of the IELTS speaking scale, the band 4 level descriptor reads, "uses a limited range of pronunciation features; attempts to control features but lapses are frequent; mispronunciations are frequent and cause some difficulty for the listener" (http://www.ielts.org/PDF/UOBDs_SpeakingFinal.pdf). Similarly, the level 2 descriptor for the TOEFL iBT "Integrated Speaking Rubrics" (Educational Testing Service, 2009, p. 190) states, "speech is clear at times, though it exhibits problems with pronunciation, intonation, or pacing and so may require significant listener effort. . . . Problems with intelligibility may obscure meaning in places (but not throughout)." These descriptors only vaguely reference the error types that lead to listener difficulty. In addition, the use of the term "pronunciation" differs across the scales. In the IELTS scale, "pronunciation" could be interpreted as referring to both segmental (individual sounds) and suprasegmental phenomena (e.g., intonation, rhythm, word stress), although this is not specified. In contrast, in the TOEFL iBT, the juxtaposition of "pronunciation" with "intonation" suggests that "pronunciation" refers only to segmental features. Clarifying the meaning of "pronunciation" is necessary to convey what exactly is being measured and is crucial for score interpretation.

Scales that employ relativistic descriptors offer even less clarity about the focal construct. For example, Morley's (1991) Speech Intelligibility Index makes reference to "basically unintelligible," "largely unintelligible," "reasonably intelligible," "largely intelligible," and "fully intelligible" speech (p. 502). However, these semantic differences do little to guide raters on how the qualities manifested in test takers' performance samples align with the scale levels.

Finally, a major shortcoming in the way that pronunciation is modeled in current L2 oral proficiency scales is that some scales conflate the dimensions of

comprehensibility and accentedness. For example, the highest level of the Cambridge ESOL "Common Scale for Speaking" groups "easily understood" pronunciation with "native-like" control of "many features" (University of Cambridge ESOL Examinations, 2008, p. 70). Similarly, the Speech Intelligibility Index systematically equates increases in comprehensibility with decreases in the interference of accent until the highest level, when "near native" speech is achieved and "accent is virtually nonexistent" (Morley, 1991, p. 502). However, a large volume of L2 pronunciation research has shown that comprehensibility and accentedness, while related, are partially independent dimensions (Derwing & Munro, 2009). That is, L2 speakers with detectable L1 accents may be perfectly understandable to their listeners, whereas speech that is difficult to understand is almost always judged as being heavily accented. Clearly, there is a need for a greater understanding of the linguistic factors that underlie L2 comprehensibility ratings, particularly at high levels of ability, so that reference to accent or native-like speech can be left aside.

Current Research

Overview

Although the increased visibility and momentum of L2 pronunciation within the broader field of applied linguistics over the past few years is evidenced in pronunciation-specific journal special issues, invited symposia, special interest groups, and, most recently, in the establishment of the annual Pronunciation in Second Language Learning and Teaching conference, this momentum has yet to extend to L2 pronunciation assessment specifically. This notwithstanding, there are two areas in the L2 assessment literature in which discussions on pronunciation are noteworthy. One is in the North American literature on international teaching assistants (ITAs) in light of concerns about ITAs' spoken proficiency; the other is in the growing body of research on automated scoring for L2 speaking—a subject that is likely to continue to inspire debate as speech recognition technologies become increasingly sophisticated and implementable in a variety of assessment contexts. Both areas will be discussed in the remainder of the chapter. In particular, research aimed at gaining a deeper understanding of major holistic constructs in L2 pronunciation research will be emphasized.

Linguistic Influences on L2 Intelligibility and Comprehensibility

In an increasingly globalized world with greater human mobility, a growing number of students face the challenge of conducting academic tasks in their L2. This includes international graduate students who bear instructional responsibilities in higher education settings in a medium of instruction that is different from their L1, referred to here as ITAs. ITAs' pronunciation has been singled out as problematic by different university stakeholders, including undergraduate students, English for academic purposes experts, and ITAs themselves (Isaacs, 2008). However, "pronunciation" (or "accent") sometimes serves as a scapegoat for

other linguistic or nonlinguistic barriers to communication that may be more difficult to identify (e.g., ITAs' acculturation issues or listeners' discriminatory attitudes toward accented speech; see Kang & Rubin, 2009). In cases where listener understanding is genuinely at stake, targeted training of the factors that are most consequential for achieving successful communication should be prioritized in ITA instruction and assessment while taking into account their teachability/learnability (e.g., for adult learners with "fossilized" pronunciation). Unless concrete, empirically substantiated guidelines on what matters most for intelligibility and comprehensibility are provided to teachers, there is a risk that pronunciation features that are perceptually salient (i.e., are noticeable or irritating) but that have little bearing on listener understanding will be targeted (e.g., English interdental fricatives) in lieu of features that have more communicative impact (Derwing & Munro, 2009).

Jenkins (2002) proposed a core set of pronunciation features that should be emphasized in instruction for a new, global variety of English—the "lingua franca core." Although her argument for a transnational standard of English that is an alternative to native speaker varieties is compelling, her recommendations are based on a limited data set. Further, the inclusion criteria for speech samples in the English as a lingua franca corpus that Jenkins and her colleagues frequently cite have not been clarified (e.g., Seidlhofer, 2010). Therefore, substantially more empirical evidence is needed before the lingua franca core can be generalized across instructional contexts or adopted as a standard for assessment.

To date, only a handful of empirical studies have examined which pronunciation features are most important for intelligibility and comprehensibility. Perhaps the most conclusive findings arise from controlled studies that have systematically isolated a particular pronunciation feature to examine its effect on intelligibility (narrowly defined; see above). Generally, different experimental conditions are created either through manipulating sound files using digital editing techniques (e.g., for syllable duration) or through having the same speaker record different renditions of an utterance (e.g., correct versus displaced primary stress placement). Taken together, the studies reveal that that prosodic (i.e., suprasegmental) aspects of pronunciation related to stress and timing have a direct effect on intelligibility (e.g., Hahn, 2004), although other features have yet to be methodically examined. This emerging evidence supports previously unsubstantiated claims about the negative effects of prosodic errors on communication.

As for segmental errors, the available evidence suggests that a nuanced approach to instruction and assessment is needed, since some segmental contrasts (e.g., /s/ vs. /ʃ/ in English) appear to be more detrimental to intelligibility and comprehensibility than others (e.g., /θ/ vs. /f/). This is dependent, in part, on the frequency of the contrast in distinguishing between lexical items (i.e., the so-called functional load principle; Munro & Derwing, 2006). It is likely that segmental errors are more problematic for learners from some L1 backgrounds than others and that the occurrence of segmental errors in conjunction with prosodic errors (e.g., word stress) can be particularly problematic (Zielinski, 2008). Overall, prosodic errors seem to be more crucial for listener understanding than segmental errors, although some segmental errors clearly lead to reduced intelligibility and comprehensibility and should be addressed (Munro & Derwing, 2006). In order

to target the problem, it is important to first diagnose whether the learner's difficulty lies in perception, production, orthographic influence (particularly in languages with poor sound–symbol correspondence), or a combination of these factors. In addition to systematically testing the perception and production of target features at the individual sound, word, and/or sentential levels, in the case of speech production, a diagnostic passage (read-aloud task crafted to elicit particular segmental or prosodic features that may not occur in natural speech) could be used in conjunction with a prompt eliciting an extemporaneous L2 speech sample (see Celce-Murcia et al., 2010).

Beyond diagnosing learner problem areas for pedagogical reasons, gaining a deeper understanding of the linguistic factors that most influence listeners' L2 comprehensibility ratings is crucial for adequately operationalizing the construct in assessment instruments. In low stakes research contexts, comprehensibility and accentedness are conventionally measured using nine-point numerical rating scales (1 = very difficult to understand, 9 = very easy to understand; 1 = very accented, 9 = not accented at all; e.g., Munro & Derwing, 2006). A minority of studies have instead used sliding scales (i.e., the rater places a cursor along a continuum to indicate his/her scoring decision) or Likert-type scales with a different number of scale levels. Such scales are appealing to L2 pronunciation researchers precisely due to their generic nature, since they can be used with L2 learners from virtually any L1 background and proficiency level. However, a caveat is that the raters receive no guidance on how to make level distinctions and, in the case of the conventionally used nine-point scales, are unlikely to converge on what the nine levels "mean" in terms of performance qualities, particularly between scalar extremes where no descriptors are provided (Isaacs & Thomson, in press). While these scales have been shown to work well for rank-ordering speakers, the lack of clarity on what is being measured at each scale level limits the precision of the instruments and raises questions about the validity of the ratings (e.g., it is unclear whether comprehensibility refers to comprehensibility of the overall message or of each individual word).

In a recent study examining the linguistic factors that underlie listeners' L2 comprehensibility ratings for the purpose of deriving a preliminary L2 comprehensibility scale for formative assessment purposes, Isaacs and Trofimovich (2012) analyzed speech samples of 40 Francophone learners of English on a picture narrative task using 19 speech measures drawn from a wide range of linguistic domains, including segmental, suprasegmental, temporal, lexicogrammatical, and discourse level measures. The speech measures were analyzed using both auditory and instrumental techniques. For example, in terms of suprasegmentals, "pitch contour" at clause boundaries was measured using listeners' perceptions of pitch patterns at the end of intonation phrases (auditory), whereas "pitch range" was measured using the pitch tracker function in the Praat speech analysis software (instrumental). The analyzed measures were then correlated with 60 raters' mean L2 comprehensibility ratings using the nine-point numerical comprehensibility scale. By bringing together statistical indices and raters' accounts of influences on their judgments, it was possible to identify a subset of measures that best distinguished between three different levels of L2 comprehensibility. Overall, lexical richness and fluency measures differentiated between low level learners,

grammatical and discourse level measures differentiated between high level learners, and word stress differentiated between learners at all levels. Such a formative assessment tool could help teachers integrate pronunciation with grammar and vocabulary teaching in communicative classrooms. However, follow-up validation studies are needed to refine the scale and clarify the range of tasks and settings that scale descriptors can be extrapolated to.

The Isaacs and Trofimovich (2012) study represents an initial step at "deconstructing" L2 comprehensibility by focusing on linguistic properties of speech. However, the scores that raters assign may also be influenced by individual differences in rater characteristics—factors that are external to the test takers' performance that is the object of the assessment. This topic is examined in the next section.

The Influence of Rater Characteristics on Their Judgments of L2 Pronunciation

A growing body of L2 speaking assessment research has examined the influence of rater background characteristics on rater processes and scoring outcomes. Research focusing on L2 pronunciation specifically is a subset of this literature. In a recent study, Isaacs and Trofimovich (2010, 2011) examined the effects of three rater cognitive variables—phonological memory, attention control, and musical ability (aptitude)—on rater judgments of L2 comprehensibility, accentedness, and fluency. The rationale was that, if individual differences in rater cognitive abilities were found to influence raters' scoring, then this could pose a threat to the validity of their ratings. There were two major findings. First, no significant effects were detected for phonological memory and attention control, which is reassuring because it removes these variables as a possible source of rater bias. Second, musical raters were overall more severe in their judgments of L2 comprehensibility and accentedness than their less musical peers. Follow-up analyses revealed that musical raters' heightened sensitivity to melodic aspects of music and speech (i.e., pitch phenomena) likely accounted for these differences. Although these findings are intriguing from a research perspective, the statistical findings were relatively weak (e.g., yielded small effect sizes) and it is unclear how *practically* significant these findings are. Further evidence is needed before recommending, for example, that raters for high stakes speaking tests need to be screened for musical ability or that a homogeneous group of raters should be sought on the basis of their musical training. Therefore, until future research suggests otherwise, language testers need not be overly concerned by these findings.

Recent L2 pronunciation research has begun to establish a link between individual differences in L2 *learners'* sociolinguistic variables, such as ethnic group affiliation and willingness to communicate, and their L2 pronunciation attainment (e.g., Gatbonton & Trofimovich, 2008). Although not examined from an assessment angle, Lindemann (2002) observed that native speakers' perceptions of how well they understood their non-native interlocutors was mediated by their attitudes toward their partners' L1 (see also Kang & Rubin, 2009). Research on motivational and attitudinal factors in relation to pronunciation assessment bears further exploration.

Rater familiarity with a particular L2 accent is often not controlled for in L2 pronunciation research, and studies that have investigated this have produced inconsistent findings. Some studies have shown that greater rater familiarity is associated with a tendency toward higher scoring and better listener understanding, although other studies have found no facilitative effects (see Carey, Mannell, & Dunn, 2011; Isaacs & Thomson, in press). At least some of the difficulties can be accounted for by the multifarious ways in which familiarity, which is sometimes framed as listener experience or expertise, is defined (e.g., in terms of amount of exposure to a particular L2 accent, ESL/EFL teaching experience, or phonetic training) and the "novice," "inexperienced," or "lay" listener comparison group is defined (Isaacs & Thomson, in press). Clearly, greater consensus on the meaning of these terms in the context of L2 pronunciation research would be desirable.

Because subjective measures of pronunciation are contingent upon both the message sender and the message receiver, the effect of rater background characteristics on the rating processes and the scores assigned is important to examine. One way of removing rater idiosyncrasies from the scoring process is through automated (i.e., machine) scoring. This subject is discussed in the next section.

Automated Scoring

Lado's (1961) concern about the reliability of subjective scoring of test takers' L2 pronunciation productions can now be addressed through an alternative that was unavailable during Lado's time—automated scoring. Because the machine scoring system (i.e., speech recognition algorithm) is trained on pooled ratings across a large cross-section of human raters, it has the effect of averaging out individual rater idiosyncrasies in a way that operational ratings of L2 speech involving two or three human raters do not. Research on Pearson's fully automated Versant English Test (previously Phonepass) has revealed high correlations between machine-generated scores and human ratings (Bernstein, Van Moere, & Cheng, 2010) and has established criterion validity with traditional large-scale L2 speaking proficiency tests (e.g., TOEFL, IELTS). While this suggests that these tests are measuring a related construct, it is unlikely that the automated system is sensitive to the same properties of speech that human raters attend to when rating, which raises questions about the validity of the assessment. In fact, studies from the speech sciences literature have demonstrated that some aspects of listeners' auditory perceptions conflict with acoustic facts obtained using instrumental measures. For example, human listeners often perceive stressed syllables to be higher than they are revealed to be in spectral analysis (Crystal, 2008). Further, because pattern matching is involved in automated scoring, controlled tasks that generate highly predictable test taker output (e.g., utterance repetition, sentence unscrambling) are much easier for automatic scoring systems to deal with than spontaneous speech arising from more communicative tasks (Xi, 2010). However, the use of such constrained tasks, which, at present, are necessary to replicate scores that human raters are likely to assign, has led to concerns about the narrowing of the construct of speaking ability. Finally, automated speaking tests may

claim to measure intelligibility in the broad sense of the term. However, much of the emphasis in the automated system is on pronunciation accuracy (e.g., of vowels and consonants). While automated feedback can inform the test user of the presence of mispronunciations, the type of mispronunciations, even if specified, will not likely all have the same impact on an interlocutor's ability to understand the utterance. Thus, the need to define the pronunciation features that most contribute to breakdowns in communication also applies to the automated scoring of speech.

Because human interlocutors involved in real-world communication are the ultimate arbiter of the qualities of speech that promote the successful exchange of information (and not machines), it is important not to lose sight of human raters as the gold standard to which automated assessments must conform. It is likely that, as speech recognition technology continues to improve, automated scoring will become increasingly prominent in the language testing research literature and testing products, albeit not to the extent that it ever supplants human ratings. There will always be constraints on what the automated system is able to do.

Challenges

This article has brought to the fore key issues in L2 pronunciation assessment. Numerous challenges have emerged thus far. Among the most salient are the need to:

- unparse the role of pronunciation (i.e., "phonology/graphology") in theoretical models of communicative competence and communicative language ability;
- discontinue the use of pronunciation item types or assessment methods that do not meet high standards of reliability and validity (e.g., paper and pencil items purportedly testing pronunciation production) or that are methodologically unsound or of questionable fairness (e.g., speech analyses for asylum purposes by authorities who know little about language or linguistics), particularly when they are being used for high stakes purposes;
- clarify the role of pronunciation within the broader construct of L2 speaking ability;
- disambiguate terms in the L2 pronunciation research literature that are not used with consistency, such as intelligibility and comprehensibility or listener (rater) expertise, experience, and familiarity;
- recognize that intelligibility (broadly defined) is the appropriate goal of L2 pronunciation instruction and assessment in the vast majority of language use contexts but needs to be more clearly understood;
- prioritize empirical studies that isolate a particular segmental or suprasegmental feature to examine measurable effects of that feature on intelligibility or comprehensibility (narrowly defined), the findings of which can then be examined in conjunction with evidence from observational studies;
- develop a greater understanding of the linguistic factors that underlie listeners' perceptions of L2 comprehensibility for the purpose of operationalizing

comprehensibility more clearly in rating scales, including without resorting to a native speaker standard;
- examine systematic sources of variance (e.g., psycholinguistic, sociolinguistic, or experience-related rater variables) that have the potential to influence ratings of L2 pronunciation but that may be extraneous to the construct being measured (i.e., are possible sources of rater bias);
- provide L2 teachers with more precise information on the error types that most contribute to communication breakdowns so that these can be targeted in L2 speaking and listening instruction and assessment;
- continue to investigate the relationship between human-mediated and machine-mediated assessments of L2 pronunciation, including the extent to which automated speech recognition can predict human scoring on more communicatively oriented tasks and the quality of the feedback delivered to test users.

While these areas, both individually and as a unit, constitute major challenges, there is one challenge that underpins all of these points and that is fundamental to propelling L2 pronunciation assessment into a post-Lado era. That is, the most significant challenge in the area of pronunciation assessment research today is to reinvigorate the conversation on L2 pronunciation in L2 assessment circles. To say that the area of L2 pronunciation assessment has been under-researched over the past several decades would be an understatement, as repercussions of the view that pronunciation is incidental to L2 learning and is unessential for communicative competence still resonate. Although, in the minds of some applied linguists, pronunciation hearkens back to tedious, mechanical drills and decontextualized discrete-point items of the past, the potential for communicatively oriented items is evident in some currently available teaching materials (e.g., Grant, 2009) if it has not yet infiltrated pronunciation assessments.

Although there is no mass reversal of the marginalization of L2 pronunciation from discussions on L2 assessment, a glimmer of hope is apparent in the publication of three articles on L2 pronunciation in the prominent assessment journal *Language Testing* since 2010 (as of the writing of this chapter). These articles, on the subjects of automated assessment and rater accent familiarity effects, are only the second, third, and fourth pronunciation-focused articles to have been published in the journal since its inception in 1984. Fruitful areas for future research are discussed in the final section of this chapter.

Future Directions

As the debate on automated scoring in relation to L2 speaking has gained increasing momentum with the recent launch of fully automated tests (e.g., the high stakes Pearson Test of English Academic or the low stakes SpeechRater, intended for TOEFL iBT training purposes), the topic of pronunciation has resurfaced in L2 assessment circles. However, this is only one area of research that merits attention. If we accept the argument that pronunciation (and, in particular, broadly defined intelligibility) needs to be assessed as part of the construct of L2 oral proficiency, then there is an urgent need to better define the constructs that we intend to

measure for assessment purposes, including filtering out accentedness from L2 proficiency scales. While accentedness is of substantive interest to L2 pronunciation researchers due to its potential to influence listeners' attitudes toward L2 speech (Kang & Rubin, 2009), intelligibility (broadly defined) is by far the more important construct for L2 pronunciation pedagogy and assessment (see above). It follows that operationalizing comprehensibility in more explicit terms in rating scales without resorting to the native speaker standard should be the focus of current L2 pronunciation scale development (Isaacs & Trofimovich, 2012). From a research perspective, this could be accomplished by triangulating statistical findings of the unique components of comprehensibility versus accentedness with raters' accounts of the linguistic aspects of the speech that they attend to when rating each construct. Drawing on Isaacs and Trofimovich's (2011) finding that musical raters, who are more attuned to certain aspects of the speech signal than their less musical counterparts, overall perceive comprehensibility and accentedness to be more independent dimensions, eliciting musicians' perceptions may be helpful in teasing these constructs apart.

One final substantive area not yet addressed in this chapter that needs to be flagged as a research priority relates to examining learners' L2 pronunciation performance on tasks that elicit a wider range of interactional patterns. Most of the pronunciation research cited above has involved native speakers' ratings of non-native speakers' performances on relatively inauthentic monologic tasks. Generally, this involves L2 learners (i.e., research participants) speaking into the microphone without the presence of an interlocutor, which does not foster genuine communication. To reflect the reality of English as a global language more closely, including the likelihood that L2 learners will need to interact not only with native speakers, but also with non-native interlocutors (depending, of course, on the context), performance on more collaborative tasks that bear greater resemblance to the real-world tasks that learners will be expected to carry out would be desirable. From an L2 assessment perspective, paired speaking tasks generally involve dyadic interactions among non-native interlocutors, although pairing procedures can be somewhat haphazard. Future research could, for example, investigate the effects of same versus different L1 group pairings on factors such as communicative efficiency and the production of target-like pronunciation.

SEE ALSO: Chapter 3, Assessing Listening; Chapter 9, Assessing Speaking; Chapter 16, Assessing Language Varieties; Chapter 63, Acoustic and Temporal Analysis for Assessing Speaking; Chapter 80, Raters and Ratings; Chapter 81, Spoken Discourse; Chapter 95, English as a Lingua Franca

References

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, England: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly*, *16*, 449–65.

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*, *27*, 355–77.

Buck, G. (1989). Written tests of pronunciation: Do they work? *ELT Journal*, *43*, 50–6.

Carey, M. D., Mannell, R. H., & Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing*, *28*, 201–19.

Celce-Murcia, M., Brinton, D. M., Goodwin, J. M., & Griner, B. (2010). *Teaching pronunciation: A course book and reference guide* (2nd ed.). Cambridge, England: Cambridge University Press.

Crystal, D. (2008). *A dictionary of linguistics and phonetics* (6th ed.). Malden, MA: Wiley-Blackwell.

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

Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching*, *42*, 1–15.

Educational Testing Service. (2009). *The official guide to the TOEFL test* (3rd ed.). New York, NY: McGraw-Hill.

Fayer, J. M., & Krasinski, E. (1987). Native and nonnative judgments of intelligibility and irritation. *Language Learning*, *37*, 313–26.

Firth, J. R. (1957). *A synopsis of linguistic theory, 1930–1955*. Oxford, England: Blackwell.

Flege, J. E., Schirru, C., & MacKay, I. R. A. (2003). Interaction between the native and second language phonetic subsystems. *Speech Communication*, *40*, 467–91.

Fraser, H. (2009). The role of "educated native speakers" in providing language analysis for the determination of the origin of asylum seekers. *International Journal of Speech Language and the Law*, *16*, 113–38.

Gatbonton, E., & Trofimovich, P. (2008). The ethnic group affiliation and L2 proficiency link: Empirical evidence. *Language Awareness*, *17*, 229–48.

Grant, L. (2009). *Well said: Pronunciation for clear communication* (3rd ed.). Boston, MA: Heinle.

Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly*, *38*, 201–33.

Isaacs, T. (2008). Towards defining a valid assessment criterion of pronunciation proficiency in non-native English speaking graduate students. *Canadian Modern Language Review*, *64*, 555–80.

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly, 10*.

Isaacs, T., & Trofimovich, P. (2010). Falling on sensitive ears? The influence of musical ability on extreme raters' judgments of L2 pronunciation. *TESOL Quarterly*, *44*, 375–86.

Isaacs, T., & Trofimovich, P. (2011). Phonological memory, attention control, and musical ability: Effects of individual differences on rater judgments of second language speech. *Applied Psycholinguistics*, *32*, 113–40.

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

Jenkins, J. (2002). A sociolinguistically based, empirically researched pronunciation syllabus for English as an international language. *Applied Linguistics*, *23*, 83–103.

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*, 441–56.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. London, England: Longman.

Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly*, *39*, 369–77.

Levis, J. M. (2006). Pronunciation and the assessment of spoken language. In R. Hughes (Ed.), *Spoken English, TESOL and applied linguistics: Challenges for theory and practice* (pp. 245–70). New York, NY: Palgrave Macmillan.

Lindemann, S. (2002). Listening with an attitude: A model of native-speaker comprehension of non-native speakers in the United States. *Language in Society*, *31*, 419–41.

Major, R. C. (2007). Identifying a foreign accent in an unfamiliar language. *Studies in Second Language Acquisition*, *29*, 539–56.

Morley, J. (1991). The pronunciation component of teaching English to speakers of other languages. *TESOL Quarterly*, *25*, 481–520.

Munro, M. J., & Derwing, T. M. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System*, *34*, 520–31.

Munro, M. J., Derwing, T. M., & Burgess, C. S. (2010). Detection of nonnative speaker status from content-masked speech. *Speech Communication*, *52*, 626–37.

North, B. (2000). *The development of a common framework scale of language proficiency*. New York, NY: Peter Lang.

Seidlhofer, B. (2010). Giving VOICE to English as a lingua franca. In R. Facchinetti, D. Crystal, & B. Seidlhofer (Eds.), *From international to local English—and back again* (pp. 147–63). Frankfurt, Germany: Peter Lang.

Spolsky, B. (1995). *Measured words: The development of objective language testing*. Oxford, England: Oxford University Press.

Trofimovich, P., & Gatbonton, E. (2006). Repetition and focus on form in processing L2 Spanish words: Implications for pronunciation instruction. *Modern Language Journal*, *90*, 519–35.

University of Cambridge ESOL Examinations. (2008). *Certificate of Proficiency in English: Handbook for teachers*. Cambridge, England: UCLES.

Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, *27*, 291–300.

Zielinski, B. W. (2008). The listener: No longer the silent partner in reduced intelligibility. *System*, *36*, 69–84.

Suggested Readings

Harding, L. (2011). *Accent and listening assessment: A validation study of the use of speakers with L2 accents on an academic English listening test*. Frankfurt, Germany: Peter Lang.

Harding, L. (2013). Pronunciation assessment. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 4708–13). Malden, MA: Wiley-Blackwell.

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

Koren, S. (1995). Foreign language pronunciation testing: A new approach. *System*, *23*, 387–400.

Lippi-Green, R. (2012). *English with an accent: Language, ideology and discrimination in the United States* (2nd ed.). London, England: Routledge.

Wennerstrom, A. (2001). *The music of everyday speech: Prosody and discourse analysis*. Oxford, England: Oxford University Press.