Watts, P., & Huensch, A. (2015). Assessing assessment: A principled revision of an in-house pronunciation diagnostic test. In J. Levis, R. Mohammed, M. Qian & Z. Zhou (Eds). *Proceedings of the 6th Pronunciation in Second Language Learning and Teaching Conference* (ISSN 2380-9566), Santa Barbara, CA (pp. 253-261). Ames, IA: Iowa State University.

ASSESSING ASSESSMENT: A PRINCIPLED REVISION OF AN IN-HOUSE PRONUNCIATION DIAGNOSTIC TEST

Patricia Watts, University of Illinois at Urbana-Champaign
Amanda Huensch, University of South Florida

In the last decade, interest in L2 pronunciation research and pedagogy has steadily gained momentum; yet, less attention has been paid to the area of assessing pronunciation either separately or as part of the larger construct of speaking ability. Isaacs' (2014) chapter on assessing pronunciation bemoans this fact while also noting how assessment should and could reflect recent advances in theory and research, such as the paradigm shift from accentedness to intelligibility (Levis, 2005) and findings related to intelligibility (e.g., Munro & Derwing, 2006).

In this paper, the researchers evaluate and revise an existing pronunciation diagnostic test based on a review of pronunciation assessment literature. Some modifications of the current test included the addition of a section testing aural perception, changing the free speech section from a self-introduction to an interview, and revising the targeted segmental features based on principled selection criteria. While the focus of this paper is on a test of English pronunciation within the context of international teaching assistant training, the authors believe the insights gained will be valuable and relevant for test development in other contexts as well as for other languages.

INTRODUCTION

A number of markers indicate sustained interest in L2 pronunciation and research over the past decade. *TESOL Quarterly* devoted an issue to the topic in 2005, the Pronunciation in Second Language Learning and Teaching (PSLLT) conference—established in 2009—continues to grow, and most recently, the inaugural issue of the *Journal of Second Language Pronunciation* is slated to debut in 2015. Despite these landmarks, the sentiment that pronunciation is overlooked still prevails—overlooked in terms of curricular focus in language programs (Derwing & Munro, 2005), teacher education and training (Breitkreutz, Derwing & Rossiter, 2001), and testing (Isaacs, 2014). With regard to testing, Isaacs (2014) notes that while the subject of L2 pronunciation teaching "conjures up images of neglect," in comparison, L2 pronunciation testing does not even have a body of literature to document its current state (p. 142). Yet, the testing and evaluation of learners' pronunciation is an integral part of the teaching and curriculum development process (Celce-Murcia, Brinton, Goodwin & Griner, 2010) and—as Isaacs states—should reflect recent advances in theory and research, such as the paradigm shift from accentedness to intelligibility (Levis, 2005) and findings from the intelligibility literature (e.g., Munro & Derwing, 2006).

The aim of this study is to review literature related to L2 pronunciation testing and the aforementioned pronunciation research areas for the purpose of evaluating and revising an existing pronunciation test used in a university-level stand-alone pronunciation course. Pronunciation tests serve one of three purposes: 1) they can be used as a diagnostic to identify the specific features with which a student needs help; 2) they can measure achievement by

determining if a pronunciation feature has been learned; or 3) they can be part of the larger construct measuring overall oral proficiency (Harding, 2012). The test in this study is a diagnostic test (purpose 1) and is used in a course for international graduate students striving for a more intelligible pronunciation in order to meet the campus oral proficiency requirements to become teaching assistants. As such, the students are motivated learners and want to make noticeable improvements over the course of the semester. The diagnostic test administered during the first week of the course provides valuable information to a) help create individualized student learning plans by providing each student with a list of pronunciation features to improve and b) help with overall curriculum planning in terms of common problem areas shaped by the class members. The relatively high stakes nature of the teaching context and the centrality of the diagnostic to the course itself motivated us to ensure the grounding of the test on current literature and research findings. Based on a review of the literature, the researchers developed the following questions related to test components, pronunciation features, and rating scales to guide the revisions of the diagnostic test:

1) What are the values and limitations of read-aloud tasks and free speech tasks?
2) Which suprasegmental and segmental features should be selected for testing?
3) How can a test address the different problem areas of speakers from multiple language backgrounds equally well?
4) Is testing aural perception an important component of diagnostic testing?
5) How can rating scales reflect the paradigm shift from nativeness to intelligibility?

In the following sections, the authors summarize relevant findings from the literature and discuss the modifications made to the pronunciation diagnostic test that resulted.

RESULTS

What are the values and limitations of read-aloud tasks and free speech tasks?

One of the hallmark features of traditional pronunciation diagnostic tests is the inclusion of a read-aloud section to test learners on a variety of potential problem areas, including segmental and suprasegmental features. Read-aloud items are typically a series of sentences or a passage filled with potential problem areas that the student reads aloud. Advantages of including read-aloud passages are many. Firstly, the use of a reading passage limits the influence of other variables such as fluency, grammatical accuracy, etc. in that students are reading rather than producing their own language (Madsen, 1983; van Weeren & Theunissen, 1987). In addition, because all students read the same passage, they provide comparable speech samples for assessment purposes (van Weeren & Theunissen, 1987). Finally, the read-aloud passages can be designed to capture and/or highlight pronunciation features that might not occur as frequently in free speech but that are known to cause difficulties for learners, such as certain consonant cluster configurations (e.g., #sC) or intonation patterns (e.g., choice questions) (Celce-Murcia et al., 2010). On the other hand, there are limitations to using read-aloud passages. One of the major limitations is that reading ability becomes an 'intervening' variable in that reading may not reflect a test taker's pronunciation in spontaneous speech (Koren, 1995). In fact, the oral reading ability of literate native speakers is not universal (Celce-Murcia et al., 2010). Finally, Celce-

Murcia et al. (2010) recommends that dialogues or other conversational texts be used rather than passages for testing suprasegmental features such as intonation and prominence.

In order to overcome some of the limitations of read-aloud passages, free speech samples may also be obtained in pronunciation diagnostic tests. These sections might include a prompt or series of prompts eliciting extemporaneous speech, such as picture story narration, role plays, or interview questions. While free speech prompts contain some of the limitations that are not an issue for read-aloud tasks (students may avoid difficult targets and rating can be influenced by other variables, such as grammatical accuracy), they allow pronunciation performance on tasks that are more reflective of real world communication. Depending on the type of free speech task included, an additional advantage can be invoking interaction, such as through the inclusion of collaborative tasks and paired speaking tasks (Isaacs, 2014; Koren, 1995). One of the main critiques of using free speech tasks is that ratings can be influenced by difficulties in other areas (fluency and grammar); however, it is the case that free speech *is* ratable and can receive a score (Buck, 1989). To avoid scores on free speech tasks being influenced by other variables such as grammatical errors and hesitation phenomena, raters can be trained to ignore fluency and accuracy errors not relevant to the elements being tested (Koren, 1995).

Before and After

Overall, the findings from our literature review and our practical experience led to the conclusion that the test should include *both* read-aloud and free speech tasks because these two types "complement each other" and confirm areas of difficulty (Celce-Murcia et al., 2010; Isaacs, 2014). Our original diagnostic test included a reading passage adapted from Celce-Murcia, Brinton, and Goodwin (1996) and a free speech section that contained a self-introduction prompt. Based on our findings, we made two major modifications (see Appendix A). Firstly, in the read-aloud section, we expanded our focus on suprasegmental features. Our original passage contained 33 suprasegmental targets (in comparison to 95 segmental targets) focused on intonation, contractions, linking, h-elision, and vowel reductions. It did not include any targets for prominence or lexical stress. In our revised test, we increased the number of suprasegmental targets to 80 (25 of which focused on prominence and lexical stress) by modifying the original passage and including a dialog. The inclusion of a dialog allowed us to more easily add targets focused on prominence and intonation (Celce-Murcia et al., 2010). Secondly, we changed our self-introduction to an interview so that it was more interactional and better reflected what was expected of our students in their teaching contexts.

Which suprasegmental and segmental features should be selected for testing? And, how can a test address the different problem areas of speakers from multiple language backgrounds equally well?

One source of information that should inform the decision about which pronunciation features to include is the literature on intelligibility and comprehensibility. *Intelligibility* is the "extent to which a native speaker understands the intended message" (Derwing & Munro, 1997, p. 2), and *comprehensibility* means "how difficult or easy an utterance is to understand" (Derwing & Munro, 1997, p. 2). Interest in the types of pronunciation errors that impact intelligibility the

most has been strong. Studies have noted the impact of both non-standard suprasegmental and segmental features to intelligibility and comprehensibility.

In terms of suprasegmentals, Munro and Derwing (1995) linked prosodic aspects of speech to raters' perceptions of comprehension. Hahn (2004) found that prosody—especially prominence at the sentence level—affected both overall comprehensibility and native speakers' reactions to the accent. Incorrect lexical stress also contributes to decreased intelligibility as has been noted by a number of scholars (Benrabah, 1997; Zielinski, 2008). In terms of testing, Koren (1995) noted the significance of stress and intonation at the phrase and word level. While Jenkins (2000) has presented evidence against including features of blended and connected speech in the lingua franca core and even in some native speaker settings, she does note three specific situations in which the addition of a full range of suprasegmentals is warranted: contexts in which 1) learners who will interact primarily with native speakers; 2) learners live in an English speaking country for extended periods; and 3) learners who want to sound native-like for professional or personal reasons (p. 136). Our specific teaching context does indeed fall within the constraints offered by Jenkins as our learners live in the US, interact extensively with native speaking students, and some—though not all—desire to sound native-like.

As for segmentals, data that would establish a rank order of segmentals according to their impact on intelligibility has not emerged, but Zielinski (2008) found that segmental errors (both consonants and vowels) in strongly stressed syllables impacted intelligibility the most. Additional information to help prioritize segmentals for the purpose of testing (and teaching) exists, however. Lado (1961)—in his seminal book on language testing which is still referenced today—advised testers to beware of targeting features that stand out as perceptually different but do not greatly influence understanding. One example of this might be a focus on /ð/ in function words, such as *this* or *that.* The functional load principle also yields valuable information to prioritize segmentals. Following from Brown (1991), functional load is a "gauge of the frequency with which two phonemes contrast in all possible environments" (p. 212). The more often phonemes contrast—for example the /l/ in *lap* and the /n/ in *nap*—the higher its functional load. The higher the functional load, the more important the phoneme is for inclusion in instruction (Brown, 1991; Munro & Derwing, 2006) and for testing (van Weeren & Theunissen, 1987). Lastly, given that most tests serve speakers from mixed language backgrounds, consideration of population-specific difficulties for relevant L1s is another appropriate means for selecting segmentals (van Weeren & Theunissin, 1987). Books—such as Avery and Ehrlich (1992) and Swan and Smith (2002)—enumerate common phonological errors by language background and can provide relevant information in this regard.

Before and After

The diagnostic changed considerably in terms of the pronunciation features tested. At the suprasegmental level, we added items to the read-aloud portion that tested lexical stress and prominence as well as expanded coverage of features related to blended and connected speech, such as vowel reduction in common function words, linking, and elision. At the segmental level, we were able to cull the number the segmentals significantly—from thirty to eighteen—by following the functional load principle and primarily focusing on segmentals that are problematic for the two most prevalent L1s in the course: Chinese and Korean. A partial list of the selected

segments follows: consonant clusters (e.g., /fl/), voicing of stop sounds, /p, f/, /r, l, n/, /w, v/, /ɛ, æ/, /ʌ, ɑ, ɔ/, and most tense-lax vowel contrasts.

Is testing aural perception an important component of diagnostic testing?

In addition to including components that test the oral production of learners, it is also necessary to consider their aural perception. Some examples of test items that evaluate the test taker's ability to perceive sounds or differences between sounds would be those that test sound to graphic symbol perception, dictation, or minimal pairs. Learners might be asked to accurately distinguish minimal pairs or indicate the syllable in a word that received lexical stress.

Perception and production of pronunciation are widely recognized as different skills (Koren, 1995; Lado, 1961). Testing listening discrimination skills is important because they are part of the development process for oral production (Celce-Murcia et al., 2010). Testing perception has the additional benefit of allowing a teacher to discover if test takers' difficulties are related to perception, production, or spelling (Isaacs, 2014). Ultimately, we wanted to develop a pronunciation diagnostic that systematically tests perception and production at sound, word, and phrase levels (Isaacs, 2014).

While there are an overwhelming number of advantages of testing aural perception skills on a pronunciation diagnostic exam, there is also a downside to consider. Namely, diagnostic tests must be logistical and practical in addition to valid and reliable (Lado, 1961). Testing perception increases exam time and grading load. Ultimately, however, we concluded that perception tasks should be included to provide a more complete picture of learner oral development given that perception is a related, but separate component of oral production (Celce-Murcia et al., 2010; Koren, 1995; Lado, 1961).

Before and After

Our original diagnostic did not include explicit perception items because we thought doing so would greatly lengthen the time needed to test each student individually. However, given the importance of testing aural perception skills, both segmental and suprasegmental sections were added, making perception a full-formed part of the test (see Appendix B). Targets were chosen based on the selection principles from the previous section. Segmental perception included ten discrimination items and ten identification items, and suprasegmental perception included ten lexical stress items, nine prominence items, and nine reduction items.

How can rating scales reflect the paradigm shift from nativeness to intelligibility?

Isaacs (2014) has noted that the paradigm shift away from accentedness and toward intelligibility should impact rating scales and the language they contain. To that end, she recommends the following:

- Explicitly defining terms raters may interpret differently—such as *pronunciation, comprehensible, intelligible*

- Including references to specific error types to ensure the scales are informed by the linguistic factors that lead to comprehensibility, for example, including information about sentence level prominence
- Avoiding relativistic descriptors, such as "basically unintelligible"

In addition to scales, another foundation of reliable rating is training and norming (Celce-Murcia et al, 2010; Isaacs, 2014). Specific advice includes the need to guide raters on how qualities manifested in test takers' performance align with scale levels (Isaacs, 2014), for example, the prominence (or lack thereof) in a speech sample and its impact on overall comprehensibility. In terms of assessing diagnostics, van Weeren and Theunissen (1987) advise instructors to keep a written record noting specific pronunciation errors while scoring. In our own experience, we have found it helpful to listen to a speaker's performance on the read-aloud section first, taking notes, and then listen to see if performance on the free speech sample confirms or disconfirms the initial judgment. Raters, especially those who are less experienced, also benefit from knowing in advance that a speech sample will require multiple listenings.

Before and After

Our original test did not include a holistic rating scale, nor did students receive a global performance rating. While it is possible to rate diagnostic samples in this manner, we chose not to in order to avoid any confusion or conflict with the holistic ratings our students received on the campus-wide ITA test—a test administered outside of our teaching context and not rated by the instructors. The importance of maintaining a united front and not sending a message which students might perceive as contradictory was paramount. In place of a holistic score, students received a checklist noting performance in key areas, a prioritized list of pronunciation features they needed to improve the most, and links to relevant information and resources to guide their study.

A rater training session using speech samples from student performance was already in place. No significant changes were made to rater training other than to update the training samples to reflect the new test.

DISCUSSION

While the original version of our diagnostic test worked effectively, it was not grounded in the research literature. The new version needed a trial run to test its functionality. Instructors using the new version during the fall 2014 semester found that it provided accurate and useful information about student performance and offered no suggestions for improvement. It should, however, be noted that the instructors were new to the teaching context and had no previous experience with the old version of the test. The concern that adding a perception section would greatly lengthen the testing time was unfounded, as all of parts of the test were easily completed within the allotted fifteen minute time slot.

Done well, diagnostic testing requires a significant investment, especially in terms of rating and providing feedback. This investment will yield the greatest return if diagnosis is seen in the larger framework of learning-oriented assessment, which aims to support learning through

feedback. To that end, providing students with diagnostic performance results, priorities for improvement, and resources for independent study sets the learning process in motion. Continued feedback and evaluation can help measure progress made on a specific pronunciation feature or alert students and teachers where additional time or effort is needed for improvement.

We have the good fortune to teach a stand-alone pronunciation course in which it makes sense to thoroughly diagnose students' pronunciation strengths and weaknesses. The test we developed is specific to our context, but we feel the advice provided here is generalizable to other contexts and languages if modified accordingly. More specifically, care needs to be taken to ensure the test fits with students' overall proficiency level and is geared toward the kinds of speaking the students will be expected to do. Even in the case of focusing on pronunciation as part of overall proficiency, Madsen (1983) recommends testing a few pronunciation targets, especially if students can be retested on those features again to measure improvement.

Based on our reading, we discovered a growing body of literature to guide the pronunciation test development/revision process, a trend that we hope continues so that a greater understanding of how pronunciation relates to overall oral proficiency and comprehensibility can be understood.

ACKNOWLEDGEMENTS

We would like to thank Sue Ingels and Veronica Sardegna for their help in creating the original diagnostic as well as Marissa Barlaz, Chelsea Coronel, and Christine Wingate for their feedback on the current version.

ABOUT THE AUTHORS

Patricia Watts is the Coordinator of the International Teaching Assistant program at the University of Illinois at Urbana-Champaign. Her areas of interest include ITA training, oral fluency, materials development, and technological applications for the teaching of pronunciation. Patricia Watts, University of Illinois at Urbana-Champaign, 4080 Foreign Languages Building, 707 S. Mathews Ave., Urbana, IL 61801, (217) 333-1506, pawatts1@illinois.edu

Amanda Huensch earned her Ph.D. and MATESL from the University of Illinois at Urbana-Champaign. She is currently Assistant Professor of Applied Linguistics at the University of South Florida. Her research interests include the acquisition of second language phonology, L2 fluency development, and ESL/TESL pedagogy. Amanda Huensch, University of South Florida, CPR 419, 4202 E. Fowler Ave., Tampa, FL, 33620, (813) 974-2548, huensch@usf.edu

REFERENCES

Avery, P., & Ehrlich, S. (1992). *Teaching American English pronunciation*. Oxford: Oxford University Press.

Benrabah, M. (1997). Word stress: A source of unintelligibility in English. *IRAL, 35,* 157-165.

Breitkreutz, J., Derwing, T. M., & Rossiter, M. J. (2001). Pronunciation teaching practices in Canada. *TESL Canada Journal, 19*, 51-61.

Brown, A. (1991). Functional load and the teaching of pronunciation. *Teaching English pronunciation: A book of readings.* New York: Routledge.

Buck, G. (1989). Written tests of pronunciation: Do they work? *ELT Journal, 43,* 50-56.

Celce-Murcia, M., Brinton, D. M., & Goodwin, J. M. (1996). *Teaching pronunciation: A reference for teachers of English to speakers of other languages.* Cambridge, UK: Cambridge University Press.

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

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

Derwing, T. M., & Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly, 39, 379-397.*

Harding, L. (2012). Pronunciation assessment. In C. Chapelle (Ed.), *Encyclopedia of Applied Linguistics* (pp. 1-6). New York: Blackwell Publishing.

Hahn, L. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly, 38,* 201-223.

Isaacs, T. (2014) Assessing pronunciation. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 140-155). West Sussex, UK: Wiley-Blackwell.

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

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

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

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

Madsen, H. S. (1983). *Techniques in testing*. Oxford, UK: Oxford University Press.

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 49,* 285-310.

Munro, M. J., & Derwing, T. M. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System, 34*, 520-531.

Swan, M., & Smith, B. (2002). *Learner English: A teacher's guide to interference and other problems.* Cambridge, UK: Cambridge University Press.

van Weeren, J., & Theunissen, T. J. J. M. (1987). Testing pronunciation: An application of generalizability theory. *Language Learning, 37,* 109-122.

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

Appendix A
Example Diagnostic Items (selected portions of read-aloud and free speech tasks)

Directions: Review the dialogue below. Rehearse by reading aloud once or twice. When you are ready, begin.

Ben: So, how was your meeting with your new advisor?
Liz: Very informative, actually
Ben: Oh, really?
Liz: There's a good probability that I'll be a TA for him next semester.
Ben: Oh, that's great. What'll you do?
…

Part 2: Free Speech Task
Example Interview Questions

1. What do you study? Why are you in that field?
2. What problems do you think you have with oral English (oral communication)?
3. What do you hope to improve this semester?

Appendix B
Example Diagnostic Items (suprasegmental perception)

Directions: Indicate the word that receives the most stress/focus in each sentence.

Example: Where are you going? Answer: ___going_____

Dialog:

A: Where are you studying? _____
B: Champaign, Illinois. _____
A: Do you like it there? _____
B: Oh, definitely. Except the winter. _____