

Evaluating Pronunciation

Constructs, Features, and Issues of Pronunciation Assessment

Julia Hilliard
17 March 2014

In the following I review current research in order to answer three research questions pertaining to assessment of the pronunciation of English language learners:

1. What are the constructs used in evaluating the productive pronunciation of English language learners?
2. Which features of English pronunciation affect these constructs? Are they all necessarily focused on pronunciation?
3. What are the practical issues associated with the assessment of pronunciation?

Annotated Bibliography

Isaacs, T. (2008). Towards defining a valid assessment criterion of pronunciation. *The Canadian Modern Language Review*, 64(4), 555-580. Retrieved from <http://muse.jhu.edu/journals/cml/summary/v064/64.4.isaacs.html>

Isaacs' study in defining a valid assessment criterion of pronunciation proficiency in International Teaching Assistants (ITA) sheds some light on the construct of *intelligibility*. The author notes that the concept of intelligibility is rather evasive, and mentions that researchers have a tendency to specialize the definition of intelligibility for the purposes of their own studies. Isaacs therefore synthesizes a few different definitions of intelligibility for her study. She draws upon Derwing and Munro's (1997) definition: the amount of the utterance that the listener understands, Ludwig's (1982) definition: the degree to which the interlocutors understands what is said, and Smith's (1992) definition: word/utterance recognition. She states that the focus of intelligibility is on the word-level, thus making it more subjective and easier to quantify.

To measure this newly synthesized construct, Isaacs adapted the ESL Placement Test's subjective word-level measure of intelligibility in order to determine whether intelligibility was an adequate assessment criterion for evaluating the pronunciation of ITAs. Given that intelligibility was deemed to be sufficient, the author also wanted to know whether a minimum level of intelligibility could be determined. For this study, Isaacs recruited 8 non-native English-speaking (NNS) participants and had each of them record 17-minute speech samples for 18 untrained native English-speaking (NS) participants/raters to listen to. To assess the intelligibility of the NNSs, the NS raters were asked to rank-order a given list of 'pronunciation problem areas' for each speaker they heard. These items included speech clarity, rate of speech, pitch contouring, sentence rhythm, word stress, and individual consonant and vowel sounds. Afterward, the NS listeners were asked to mark the approximate percent of words that they were able to understand. Finally, the NSs were asked whether or not they believed that the NNS they listened to had adequate enough pronunciation to be a TA for an undergraduate course.

Annotated Bibliography

The NNSs who received high ratings in intelligibility also received similarly high ratings in comprehensibility. And these same NNSs were also judged to have adequate pronunciation to be a TA for an undergraduate course. The results of this study indicated that intelligibility, as was defined for this study, was an adequate assessment criterion for these NNSs in the academic domain.

This study has demonstrated that intelligibility is a construct used to assess pronunciation based on word-level features. It may have different definitions in different research, but in the case of this study the construct of intelligibility is fairly easy to operationalize, given that it is construed as the clarity of individual words in an utterance.

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

In this research the authors cite Munro and Derwing's (1999) work in defining the construct of *comprehensibility* as the listeners' perceptions of how easily they understand the speech of NNSs. In this study the authors wish to determine which linguistic (and nonlinguistic) features listeners attend to when judging the comprehensibility of non-native (NN) speech.

The researchers enlisted 40 NNSs who all shared a French L1 to participate in this study. Each participant was recorded describing an eight-panel picture story, with each recording having a mean length of approximately 30 seconds. The raters in this study were 60 NS undergraduate students majoring in different nonlinguistic disciplines, all of which reported having grown up in monolingual homes. Each of the recordings were rated by these listeners according to measures in phonology—segmentals and suprasegmentals, fluency—temporal measures and number of pauses, linguistic resources—grammatical and lexical measures, and discourse—storytelling strategies and use of cohesive devices. The authors hypothesized that the variables affecting listeners' perceptions of comprehensibility may not be restricted to the

Annotated Bibliography

domain of productive pronunciation; they may also be influenced by the speakers' use of diverse grammatical or lexical content, or the speakers' overall skills as orators.

The results of the study showed that the comprehensibility judgments were indeed not completely centered in productive pronunciation; lexical measure such as the total number of unique words per unit of time, grammatical accuracy, and story breadth were also proven to be factors in NSs' judgments of comprehensibility for these NNSs. However, this was in addition to the comparably more salient measure—presence or absence of primary word stress in polysyllabic words, which was a feature of productive pronunciation. In other words, NS listeners' judgments of comprehensibility still largely relied on productive pronunciation. However, the listeners' additional consideration of grammatical, lexical, and discourse features should still be equally regarded as features that affect listeners' perceptions of this apparently pronunciation-oriented construct.

This study introduced the construct of comprehensibility, and offered set of linguistic criteria by which some NSs may assess the comprehensibility of NNSs. This information shed some light on what it might mean to be comprehensible by NS standards. However, there is more to be said about the variables that can influence listeners' perceptions of the comprehensibility of NN speech, as will be seen in the following sections.

Kang, O., Rubin, D., & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. *The Modern Language Journal*, 94(4), 554-566. doi:10.1111/j.1540-4781.2010.01091.x

Although the title of this research implies that the study is assessing the construct of *accentedness*, a dated concept in the field of pronunciation assessment relating to the perceptual degree to which a NNS's speech differs from that of a NS, this study is actually assessing the construct of *comprehensibility*, which will be clearer once the method of this study is introduced. In this study, the authors explore the relationship between NSs' comprehensibility judgments of

Annotated Bibliography

NN speech and the use of suprasegmental (prosodic) features. These features include speaking rate, pause structure, nonstandard word stress, pitch range, and intonation structure.

The data used in this study consisted of responses recorded by 26 NNS participant to an iBT TOEFL integrated task in which they were asked to summarize and demonstrate their understanding of a passage that they had just read. The responses were recorded in minute-long samples and then transcribed by the researchers in a manner that allowed them to objectively measure the presence and relative frequency of the aforementioned prosodic features in each sample. They then had 188 NS undergraduate students listen to the NNSs' speech samples and rate their ability to comprehend the content of the recording using 5 different 7-point bipolar scales: 1)easy/hard to understand, 2)incomprehensible/highly comprehensible, 3)needed a little effort/lots of effort to understand, 4)unclear/clear, and 5)simple/difficult to grasp the meaning. After collecting the NSs' ratings for each of the samples, they were then compared with the occurrence of prosodic features that were objectively measured for each NN speech sample prior to the NS ratings being taken.

The results indicated that 50% of the judgments made by NS raters were attributable to the NNSs' performance in using (or not using) the aforementioned prosodic features. The most significant features predicting NSs' of comprehensibility judgments were the ones related to fluency, or in this case, speaking rate. This finding strengthens the findings of the previous study which also suggested that features related to fluency are influential in determining NSs' comprehensibility judgments NNSs.

Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly*, 38(2), 201-223. Retrieved from <http://ejournals.ebsco.com/direct.asp?ArticleID=43A781BE55EC10A12F2F>

For the sake of remaining consistent with the currently established definitions of comprehensibility and intelligibility, I am referring to the construct that is measured in this study

Annotated Bibliography

as comprehensibility rather than the titular construct, intelligibility. This is because Hahn's definition of intelligibility, which is made apparent in the sections describing the measures on pp. 207-208, encompasses the raters' perceived ability to understand what the speaker is getting at. This definition is congruous with the concept of comprehensibility mentioned in the research reviewed thus far. It has been noted by several researchers that the concepts of intelligibility and comprehensibility are very often used interchangeably, which contributes to their ambiguity within the field of pronunciation assessment. It should be noted, however, that here there will be a distinction made between these two concepts, which is essentially that intelligibility can be quantified simply by having raters attempt to transcribe the words that they heard the speaker say. Comprehensibility, on the other hand, generally requires raters to approximate or demonstrate their comprehension of the meaning of what they heard the speaker say.

Following the theme of suprasegmentals, the current study treats the impact of NNSs' misuse of suprasegmentals on NS listeners' judgments of comprehensibility. Hahn begins by explaining the *given-new stress connection* (GNSC), which states that new and/or contrastive information in a sentence is presented in stressed elements, while old and/or given information remains unstressed. Research has shown that NNSs from many different linguistic backgrounds have shown a tendency to make two mistakes with the GNSC. First is misplaced primary stress in the new/contrastive elements, and second is the stressing of all words in the utterance more or less equally.

The materials of this study consisted of three speech recordings of the same NNS. These recordings were three versions of the same academic lecture which were adapted by the author for the purposes of this study. The versions were identical except that Version A did not violate the GNSC, while Versions B and C violated the GNSC in different ways. Version B misplaced

Annotated Bibliography

the primary stress, and Version C contained no primary stress whatsoever. The participants in this study were 90 NS undergraduates coming from homes where only English was spoken. These participants were asked to give three evaluations based on the recordings that they heard. The first of which measured their difficulty in processing the discourse, the second measured their overall comprehension of the recording, and the third measure their subjective reactions to the speaker and the recording.

The results showed that the NSs generally responded more positively to the recording of Version A, which did not violate the GNSC. They displayed an ability to quickly process the information in the recording, scored better on the comprehension tasks, and gave more positive attitudinal ratings of the NNS recorded.

This study strengthens the findings of previous studies linking the successful use of suprasegmentals to the prediction of NSs' judgments of the comprehensibility of NN speech.

Kennedy, S., & Trofimovich, P. (2008). Intelligibility, comprehensibility, and accentedness of L2 speech: The role of listener experience and semantic context. *Canadian Modern Language Review*, 64(3), 459-489. doi:10.3138/cmlr.64.3.459

From this point on, the focus will be shifting to the listeners of the NNSs. In the field of pronunciation testing, the true value of a test score is determined by what it predicts about a given speaker's communicative success outside of the classroom. That said, the focus will be shifting to the listeners of the NNSs. The current research considers the degree to which NSs' experience with/exposure to NN speech may affect their intelligibility, comprehensibility, and accentedness judgments of NN speech.

The two variables in this study were the NSs' differing amounts of exposure to NN speech in their lifetimes, and their ratings of intelligibility, comprehensibility, and accentedness of recorded utterances containing varying degrees and types of semantic information. The use of varying degrees of semantic information in the recorded utterances in this study allowed the

Annotated Bibliography

researchers to isolate the variables and to more objectively measure the NSs' ability to comprehend and transcribe the recorded utterance.

The participants who provided the recordings were 6 NNSs and 6 NSs. The materials consisted of 3 lists of sentences. List 1 was a series of single-clause sentences that all contained a real-world context, semantically speaking. These sentences were to be marked as True or False based on the listener's knowledge of the world. Lists 2 and 3 contained both semantically meaningful and semantically anomalous sentences—all of which were grammatically correct.

The group of NSs participating as listeners in this study consisted of 12 ESL teachers with at least 3 years of teaching experience and 12 undergraduate students who reported having very little exposure to NN speech. To measure their intelligibility ratings of List 1, these listeners were asked to transcribe each of the sentences after listening to them. To measure their ability to comprehend the meaning of List 1, the listeners were then asked to mark the sentences they had just transcribed as either True or False statements. Afterwards, they were asked to make two ratings for every sentence in Lists 2 and 3 using 2 Likert Scales based on their perceived accentedness and comprehensibility, respectively.

The results of this study indicated that the presence of semantic context affected ratings of intelligibility, comprehensibility and accentedness of the NNSs' recordings more so for the monolingual undergraduates than for the ESL teachers. These results reveal a rather significant limitation of pronunciation assessment; the teacher-raters' greater exposure to NN speech than most other NSs outside of the classroom will result in test-takers receiving feedback that will not be applicable to the world outside of the classroom. Teachers and test writers will need to consider their specialized ability to comprehend NN speech when developing materials for pronunciation assessment in order to give the test-takers more practically valuable feedback.

Annotated Bibliography

Lindemann, S. (2010). Who's "unintelligible"? The perceiver's role. *Issues in Applied Linguistics*, 18(2), 223-232. Retrieved from <http://escholarship.org/uc/item/89f0w1ch>

Lindemann's research brings another dimension to the listener-rater issue, focusing more on the types of listeners that NNSs will inevitably need to communicate with at some point outside of the classroom setting. In the present study, the author sought to find out whether listeners' attitudes toward the ethnicities of certain nonnative speakers' of English affected those listeners' judgments of comprehensibility and accentedness of certain NNSs.

The NS participants were selected based on their negative or positive ratings of pre-recorded speech samples of the NNS participants regarding their accentedness, ability to sound intelligent, and how 'nice-sounding' they were. A total of 12 NS listeners were selected based on this pretest; 6 positive listeners and 6 negative listeners. The NNS participants in this study were 6 NNS students all sharing a Korean L1. The study required that each NNS perform a one-on-one information gap activity with a NS. Each NNS participant partook in the activity two times; once with a negative NS participant and once with a positive NS participant.

The information gap activity was a map task in which each participant was given a map of the same imaginary place. The NNS's map had a route drawn on it, while the NS's map lacked a route. It was the NNS's job to tell the NS how to draw the same route on their own map without showing each other their maps or visually representing the shape of the route (through drawing the route in the air, for example). The NNSs were given a different map for each time they partook in the activity.

The results show that 10 out of the 12 interactions resulted in the proper route successfully being drawn on the NS's map. The two unsuccessful interactions were both with negative NS participants. It was also noted that two of the successful negative NS participants dominated the interaction with the NNS. Throughout the activity, these NSs were not

Annotated Bibliography

acknowledging the NNS's role as the more-knowledgeable other and as the main provider of input during the interaction.

This study indicated that some NSs may perceive some NN speech to be accented and incomprehensible simply due to their personal biases toward the speaker's ethnicity. This may cause some teachers/test-givers to consider the practicality of “err[ing] on the conservative side with consideration of the ‘lay’ listeners whom the student will meet” (Morley, 1994; cited in Isaacs, 2008). Perhaps this is not the way to raise students’ awareness of these particular biases in some NSs. Regardless, chances are that students will end up having to talk to people with these ethnic biases who will underrate students’ abilities to speak intelligibly, which can negatively affect the success of the interaction. If students are able to perform well when assessed on their pronunciation, only to run into these types of NSs and fail miserably when trying to interact, then that is certainly an issue that teachers will need to consider addressing in the classroom, since there is no fair way to incorporate this type of experience into assessment.

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

The researchers in this study shifted the perspective a bit to address the possibility of differing perceptions of NNS intelligibility, comprehensibility, and accentedness between groups of listeners from different linguistic backgrounds.

In the current study, the authors aim to determine whether listeners’ judgments of intelligibility, comprehensibility, and accentedness of NN speech will compare between groups of 1) NN listeners from different L1 backgrounds and 2) NN listeners and native (N) listeners.

To determine this, the authors recruited 48 NNSs from 4 different L1 backgrounds—12 each from Cantonese, Japanese, Polish, and Spanish L1 backgrounds—to each provide a short, impromptu narrative by describing a cartoon story that they had viewed prior to recording the

Annotated Bibliography

narrative. The participants that were recruited to be listeners were 40 in total; 10 NS, 10 NNSs sharing a Japanese L1, 10 NNSs sharing a Cantonese L1, 10 NNSs sharing a Mandarin L1.

To measure listeners' intelligibility ratings, they were asked to transcribe all of the approximately 7 second-long recordings after listening to each of them. Afterwards, they were asked to rate each speaker's comprehensibility on a 9-point Likert Scale, and then rate their accentedness on a similar scale.

The results showed that for both NS listeners and NNS listeners there was a moderately high degree of agreement between each listener's judgment of intelligibility, comprehensibility, and accentedness. This study's findings are important to keep in mind in light of the potential issues stemming from differences between listeners that were found in the previous reviewed research. The current research, however, demonstrates the fact that differences between listeners will not always cause unreliable variation in ratings.

Sewell, A. (2013). Language testing and international intelligibility: A Hong Kong case study.
Language Assessment Quarterly, 10(4), 423-443. doi:10.1080/15434303.2013.824974

For the current study, the perspective will be shifting once more in terms of the setting in which the NNS's interactions take place. Up until this point, the studies that have been reviewed were for ESL settings. This study will be providing some insight into a more global perspective in the field of pronunciation assessment.

In this study, the author draws upon Jenkins' (2000) criteria the Lingua Franca Core (LFC) of features that play an important role in predicting NNSs' intelligibility in international communicative contexts. The LFC has been tested in international contexts between NNSs of different language backgrounds, but it has been hypothesized that these features can also be important in interactions between NSs and NNSs. Were this hypothesis confirmed to be true, it

Annotated Bibliography

would strengthen the findings of Munro, Derwing, and Morton's (2006) research, which indicated congruencies in N and NN judgments of what intelligible English sounds like. The LFC places importance on all consonant sounds (with the exception of θ and ð), allows for the simplification of consonant clusters (so long as they do not occur in word-initial position), emphasizes vowel length (variations in vowel quality is permissible), and places importance on the presence of nuclear stress within a sentence in order to separate thought groups and orient the listener to important information. What sets the LFC apart from what has been found in the research up until now is its downplaying of suprasegmental features (with the exception of nuclear stress), which previous research has found to be a significant factor in predicting NSs' intelligibility and comprehensibility ratings of NN speech.

Sewell compares the rater comments made on the spoken component of Hong Kong's Language Proficiency Assessment for Teachers of English (LPATE) to the criteria listed in the LFC. The LPATE is one of the more globally-oriented language tests, having been developed in an area of the world like Hong Kong—one of the world's centers of international commerce—where English is spoken as the international business language on a regular basis. To make this comparison, the comments were analyzed and separated into 5 categories as they related to the LFC; vowels, consonants, consonant clusters, word stress, and connected speech/intonation. The priorities of the raters were determined by the number of comments per category and the word counts of the respective comments.

Results indicated that between segmentals and suprasegmentals the ratio was a fairly balanced 60%-40%. It was noted that the balance between comments segmental and suprasegmental was not congruous with the LFC, which emphasizes the importance of segmentals over suprasegmentals. Within these two larger categories there were two

Annotated Bibliography

observations that also did not align well with the LFC: comments on consonants were more salient than vowels, and connected speech/intonation was mentioned more than word stress.

This study has provided some insight into what it means to be intelligible and comprehensible in the context of communicating in English as an International Language (EIL). There appears to be a greater emphasis placed on segmentals which was not highlighted in the previous studies based in inner-circle settings. However, it appears that suprasegmentals are still important for intelligibility in EIL settings. But on the whole, it seems to be the case that the assessment of intelligibility and comprehensibility would take a slightly different form in EIL contexts than it would in ESL contexts.