Abstract:

Intelligibility has been widely regarded as an appropriate goal for second language pronunciation teaching. Yet there is no universally accepted definition of intelligibility, nor any field-wide consensus on the best way to measure it. Further, there is little empirical evidence to suggest which pronunciation features are most crucial for intelligibility to guide teachers in their instructional choices. This mixed methods study examines whether intelligibility is an appropriate criterion for assessing pronunciation proficiency in the academic domain. Speech samples of eight non-native graduate students were elicited using the Test of Spoken English, a standardized test often used to screen international teaching assistants (ITAs). Results of a fined-grained analysis of the speech samples coupled with intelligibility ratings of 18 undergraduate science students suggest that intelligibility, though an adequate assessment criterion, is a necessary but not a sufficient condition for graduate students to instruct undergraduate courses as teaching assistants.

Keywords: intelligibility, pronunciation, International Teaching Assistants, assessment, higher education

Résumé

L'intelligibilité est communément considérée comme un objectif approprié pour l'enseignement de la prononciation en anglais comme langue seconde. Cependant, il n'existe pas de définition universelle de l'intelligibilité, ni de consensus sur la meilleure façon de la mesurer. En outre, il n'existe que peu d'évidence empirique sur les caractéristiques de prononciation essentielles à l'intelligibilité. Ce manque limite les professeurs dans leurs choix pédagogiques.

La présente étude d'une méthodologie mixte cherche à savoir si l'intelligibilité peut servir de critère d'évaluation adéquat en matière de prononciation dans le milieu académique. Des d'échantillons linguistiques de huit locuteurs non natifs du deuxième cycle universitaire ont été obtenues du *Test of Spoken English*, un examen standardisé qui est communément utilisé dans le choix des assistants internationaux à l'enseignement (AIE). Les résultats d'une analyse très pointue des échantillons linguistiques ainsi qu'une compilation d'indices d'intelligibilité provenant de 18 étudiants du premier cycle en sciences révèlent que l'intelligibilité est une condition nécessaire mais pas suffisante pour permettre aux étudiants du deuxième cycle d'enseigner à titre d'AIE au premier cycle.

Mots clés: intelligibilité, prononciation, assistants internationaux à l'enseignement, évaluation, enseignement supérieur

Introduction

Intelligibility has been widely regarded as an appropriate goal for second language

pronunciation teaching. Yet there is no universally accepted definition of intelligibility, nor a

field-wide consensus on how to best measure it (Munro & Derwing, 1999). For instance, what

some researchers call 'intelligibility' is for others 'comprehensibility' (e.g., Gass & Varonis,

1984) and the notion of 'irritation' with non-native speech is inherent in some conceptions of the

term but not others (e.g., Ludwig, 1982). In addition to these inconsistencies, there is little

empirical evidence to suggest which pronunciation features are most crucial for intelligibility to

guide teachers in their instructional choices. Only recently have studies begun to systematically

isolate features of discourse and pronunciation in order to gauge their relationship with

intelligibility in controlled settings (e.g., Hahn, 2004). Other studies have investigated

intelligibility in dyadic non-native speaker (NNS) interactions (e.g., Jenkins, 2000; 2002),

although more research is needed to generalize findings to other contexts. In short, despite the

widespread use of the term, intelligibility is an evasive concept that we know little about.

The first objective of this paper is to synthesize the ways that 'intelligibility' has been

defined and measured in the pronunciation literature. Such a synthesis may help clarify important

concepts in the field that have often been confounded as a result of differences in terminology.

The second objective of this paper is to attempt to empirically validate one measure of

intelligibility at the word level, a locus of pronunciation difficulties for many NNSs (Field,

2005), particularly in the academic domain. The overall goal of the study is, therefore, to

determine whether intelligibility yields an adequate assessment criterion of pronunciation

proficiency for NNS graduate students.

Defining intelligibility

As early as 1949, Abercrombie defined comfortably intelligible as 'a pronunciation which

can be understood with little or no conscious effort on the part of the listener' (p. 120). Several

pronunciation proponents have reaffirmed his view that 'language learners need no more than a

comfortably intelligible pronunciation' (e.g., Dalton & Seidlhofer, 1994). Kenworthy described

comfortably intelligible as 'being understood by a listener at a given time in a given situation'

and equated it with 'understandability' (1987, p. 13). In an operational definition, she elaborated

that if a NNS substitutes a certain sound for another and the listener hears a different word or

phrase than the speaker had intended to say, the result is unintelligibility. Thus, the more words

the listener is able to accurately identify, the more intelligible the speaker is. Although

Kenworthy claims that 'intelligibility has as much to do with the listener as with the speaker' (p.

14), her notion of comfortably intelligible focuses mostly on the listener. If the NNS pronounces

such that the listener constantly needs to ask for repetition or clarification – that is, if the act of

listening becomes too laborious – then the listener becomes frustrated or irritated. Being

comfortably intelligible has to do with efficiency, then, where the listener can understand the

speaker without difficulty or recourse to repetition.

Morley (1994) also makes use of the term 'comfortably intelligible' in addition to

'functional intelligibility,' 'overall intelligibility,' and just plain 'intelligibility' (i.e., without any

qualifier). While she does not define 'comfortably intelligible,' she does contend that unless

NNSs are comfortably intelligible, they often avoid spoken interaction. 'Functional

intelligibility' is similar to Kenworthy's (1987) 'comfortably intelligible' in terms of ease of

understanding a NNS, although speech that is too difficult to understand will apparently be

'irritating' for Kenworthy's listener and 'distracting' for Morley's listener (1994).

Both Morley and Kenworthy (1987) make explicit the link between intelligibility and the broader goal of communication. Morley's *Speech intelligibility/communicability index* is a 6-level framework designed to assess 'overall intelligibility' (1994). In the 'intelligibility' column, Morley describes speech in terms of listener effort and distracting features (underscored by her conception of 'functional intelligibility'). In the 'impact on communication' column, speech is correspondingly described in terms of the degree of interference of accent in getting the message across. 'Communicative thresholds' are placed before scale bands 3 and 5, which implies that intelligibility assessed at band 3 and higher is sufficient, in Morley's terms, for communication.

The idea of a 'threshold of intelligibility' is not new. Catford coined the term in 1950 and Gimson (1980) speaks of 'minimal general intelligibility' or the lowest requirement for efficiently conveying a message from a native listener's standpoint. Yet, the rationale for Morley's placement of the thresholds in her index is not clear (1994). If her rating scale is to be widely adopted or used at a particular institution for a specific purpose, empirical validation using speech samples from the appropriate population would be desirable. For instance, Morley links intelligibility with communication in terms of accent and its effect on listener perception. Yet Derwing and Munro (1997) have shown empirically that what is unintelligible is almost always judged to be heavily accented, whereas the opposite is not necessarily the case (i.e., what is heavily accented may or may not be unintelligible). They argue for 'the need to disassociate accent ratings and intelligibility in language assessment instruments, which often confound the two dimensions' (pp. 11-12). This constitutes evidence that the scale descriptors in Morley's (1994) index are not grouped together appropriately and would benefit from empirical exploration.

Derwing and Munro (1997) define 'intelligibility' as the amount of utterance that the listener understands. This contrasts with 'comprehensibility' and 'accentedness,' which are based on listener perceptions of ease of understanding non-native speech and the extent to which the speech differs from the native speaker (NS) norm, respectively. Thus, 'intelligibility,' which is the most objective of their three measures, does not contain the listener effort or irritation latent in 'comfortably intelligible' (Abercrombie, 1949; Kenworthy, 1987) or 'functional intelligibility' (Morley, 1994). To add to the definitional confusion, Derwing and Munro's 'intelligibility' (1997) is essentially equivalent to Gass and Varonis's 'comprehensibility' (1984) in the way that the constructs are operationalized. It is also similar to the objective interpretation of Ludwig's (1982) 'comprehensibility,' defined as 'the degree to which the interlocutor understands what is said or written' (p. 275) and to Smith's 'intelligibility,' defined as 'word/utterance recognition' (1992, p. 76). Ludwig, in fact, distinguishes between 'comprehensibility' and 'irritation,' noting that 'while comprehensibility can be rated fairly objectively, irritation cannot' (p. 275). So too is Derwing and Munro's (1997) 'intelligibility' more quantifiable than the various interpretations of 'comfortably intelligible.'

Measuring intelligibility

There is no universal consensus on a definition of intelligibility. Perhaps it is logical, then, that there is also no 'universally accepted way' of measuring intelligibility (Munro & Derwing, 1999, p. 289). Several studies have attempted to measure intelligibility using listeners' orthographic transcriptions. Smith (1992) measured intelligibility using fixed cloze tests. Intelligibility scores were calculated by tabulating the number of blanks that the listener was able to fill in. Gass and Varonis (1984) measured 'comprehensibility' (equivalent to Derwing & Munro's 'intelligibility') by having NSs transcribe sentences read aloud by NNSs. Scores were assigned based on discrepancies between the transcribed sentences and the story scripts from

which the sentences were drawn. A similar procedure was followed for measuring intelligibility in Derwing and Munro (1997) and Munro, Derwing, and Morton (2006).

Intelligibility has also been measured subjectively or impressionistically. In a study by Fayer and Krasinski (1987), raters assessed 'overall intelligibility' on a 5-point scale. Similarly, Anderson-Hsieh, Johnson, and Koehler (1992) elicited 'ratings of pronunciation' on a 7-point scale, with the lowest point representing 'heavily accented speech that was unintelligible' and the highest point 'near native-like speech' (p. 538). In both studies, intelligibility ratings were followed by a second listening, where raters made scalar judgements on the presence of prosodic variables in the speech (e.g., intonation). The assumption here was that intelligibility is a scalar construct (i.e., it exists on a continuum) rather than a binary (i.e., all-or-nothing) phenomenon.

Given that there is no universally accepted definition and no single reliable measure of intelligibility, the different measures that do exist need to be empirically validated for their use in different settings. The present study attempts to validate a subjective measure of intelligibility in a context where the stakes for intelligibility are thought to be high – the International Teaching Assistant (ITA) context.

Intelligibility and the ITA context

As concerns regarding ITAs' spoken proficiency and cross-cultural awareness have increased, ITAs have been written about extensively in the applied linguistics literature (e.g., Saif, 2002; 2006; Pickering, 2001). ITA populations have also been featured in intelligibility studies (e.g., Hahn, 2004; Anderson-Hsieh et al., 1992). Although pronunciation is only one of a range of factors that can affect ITAs' ability to carry out their instructional duties, poor pronunciation is the most overtly associated problem with ITAs as identified by undergraduate students, language researchers (Hoekje & Williams, 1992), and ITAs themselves (Cheng, Myles, & Curtis, 2004). ITA screening and preparatory courses are a way for institutions to ensure that

ITAs do meet oral proficiency standards. Yet screening practices at North American academic institutions vary widely, from institutionally-developed tests to no testing at all.

The *Test of Spoken English* (*TSE*) and its retired version, the *SPEAK* test, are widely used to screen ITAs (Celce-Murcia, Brinton, & Goodwin, 1996). As a measure of global speaking performance, the rating scale of the 1995 version of the *TSE* cites 'pronunciation' as one of many features to consider in rating but leaves this construct largely unanalyzed (Educational Testing Service, 1995). While the introduction of the internet-based *Test of English as Foreign Language* (*TOEFL*) in 2005 contains a speaking component, this does not reduce the need for additional oral assessment. The objectives of such an assessment would be to ensure, first, that NNS university students have the oral language skills (and concomitant intelligibility) to carry out academic tasks and, second, that ITAs' spoken language (and intelligibility) does not preclude their ability to carry out instructional duties. Therefore, describing and validating a measure of intelligibility that can accomplish both these objectives is a research priority.

An example of one measure of intelligibility, specific to an ITA context, comes from the *ESL Placement Test* (*EPT*).[1] Developed at a Midwestern university which employs a large number of ITAs, this test includes a subjective, word-based definition of intelligibility to assess both NNS students and prospective ITAs. In the first part of the oral interview, which adopts intelligibility as its central measure, the interviewer assesses 3 minutes of the candidate's unrehearsed speech on the basis of (1) being able to understand every word that the interviewee says, where 'understand' is taken to mean that each word is understood immediately without guessing or requiring additional context, and (2) determining if there is any evidence that the interviewee misunderstood anything that the interviewer said (e.g., as signalled by nonverbal cues). If no misunderstanding is detected, the interview is discontinued and the interviewee is exempted from pronunciation work. If the interviewer does detect a misunderstanding from one

direction or another, then the second part of the interview, which adopts accuracy as its central measure, is administered to determine how well the interviewee controls the content of the ESL oral course that he/she will be required to take.

What is novel about this measure of intelligibility is its focus on the word level, a locus of pronunciation difficulties for many NNSs (Field, 2005), and the relative ease with which it can be administered (cf. Turner & Upshur, 2002). Taken together, these qualities make this easy-to-administer word-based measure of intelligibility an appealing assessment tool, applicable in a variety of ITA contexts. However, prior to widely adopting this measure for assessment purposes, it is important to test and validate it in a context other than that in which it was created, with the larger goal of determining whether it yields an adequate assessment criterion in other international university student and ITA contexts.

Method

Research questions

In this study, a subjective, word-level measure of intelligibility was adapted from the *EPT* to assess the intelligibility of NNS graduate students. The primary purpose of the study is to investigate whether intelligibility is 'enough,' that is, a sufficient goal and an adequate assessment criterion, for evaluating these students' pronunciation proficiency in the academic domain. If intelligibility is deemed to be 'enough,' then is there a threshold (i.e., minimum) level of intelligibility that can be identified? If not, then what criterion might be more suitable?

Participants

The eight NNS participants in this study (4 males, 4 females) were full-time graduate students in the Faculty of Education at a Canadian English-medium university, ranging in age from 24 to 35 years ($M = 28$). They had all been schooled in a language other than English prior to moving to Canada an average of 2.9 years (range = 1-6) earlier to pursue postsecondary

degrees. The sample included two Korean speakers and one speaker each of Mandarin, Japanese, Javanese, Spanish, Serbo-Croatian, and French. The diversity of L1 backgrounds simulates the variety of accents that undergraduate students might be exposed to in real-world ITA contexts. Two speakers were employed as ITAs during the time of the data collection and one was subsequently hired as a course instructor. In addition, 18 NS undergraduate science students, ranging in age from 18 to 24 years ($M = 20$), participated in the study as untrained raters. All raters reported having normal hearing and none had had training in linguistics or language teaching.

Instruments

A 1995 version of the *TSE* was used to elicit the speech data. This standardized oral proficiency test is a semi-direct test (i.e., uses recorded and printed stimuli) that generates timed responses to items, and, therefore, might not authentically simulate real life ITA situations. However, its widespread use on North American campuses strengthens the link between the present research context and real world ITA assessment practices, thus enhancing this study's ecological validity. The development of the rating instrument was informed by the author's prior analysis of the speech data, then piloted and refined based on feedback from an undergraduate science student. The first section of the rating instrument can be seen in Appendix 1.

Procedure

Speech recording, transcription and coding

Speech recording sessions were conducted one-on-one in a quiet office. Seventeen minute speech samples, elicited using the 1995 version of the *TSE*, were recorded at 22,050 Hz and were later normalized for peak intensity. All speakers filled out consent forms and background questionnaires in accordance with ethical procedures.

The analysis of the speech data was a bottom-up process that constituted fine-grained

transcriptions of the speech samples and culminated in a holistic 'intelligibility profile' for each

speaker.[2] In the analyzed *TSE* task, the speakers assumed the role of president of a photography

club and informed club members about changes to the exhibition schedule (90 seconds).[3] The

speech data were first transcribed into the International Phonetic Alphabet (approximately 255

syllables per person). Following Munro and Derwing (1998), speaking rates (which include all

utterances and pauses) were calculated to the nearest millisecond in addition to articulation rates

(which include all utterances but exclude pauses). A transcription system for suprasegmentals

was then developed, evolving as features 'emerged from the data' (see Appendix 2 for a listing

of transcription categories). The decision as to whether or not the speech was intelligible was

suspended until all transcriptions had been completed. Color-coding was then used to identify

instances of 'unusual pronunciation that does not affect intelligibility' (e.g., phonetic errors) and

'unintelligible pronunciation' (e.g., phonemic errors with a high functional load). Finally, the

color-coded data were quantified by calculating the ratio of intelligible words (including

unusually pronounced words) to total words for each speaker. This measure excluded fillers and

counted instances of immediate word repetition as only one word.

Stimulus preparation and rating sessions

The rating session, which was set at 1 hour, included a teaching and a practice component.

To minimize potential order effects, a randomized CD was prepared for four different rating

sessions. In each session, the raters listened to a given speaker's performance on two 60-second

TSE tasks, at the end of which they were to mark the approximate percent of words that they

were able to understand on a 0-100 scale (reproduced in Appendix 1).[4] The first *TSE* task was a

six-frame picture narrative of a holidaying couple purchasing and developing film; the second

task elicited advantages and disadvantages of taking photographs on a trip versus making written

notes. After a second listening, the raters rank ordered the pronunciation problem areas, if any,

that had hindered their understanding of words, as guided by the rating instrument. The problem

areas that they could select from in their rankings included speech clarity, rate of speech, pitch,

sentence rhythm, word stress, and individual consonant/vowel sounds.

To provide a check on consistency for intelligibility ratings, the raters also assessed each

speaker for ease of understanding on a 4-point Likert scale. Unlike intelligibility, this measure

(henceforth referred to as 'comprehensibility') contains no word-level focus.[5] The raters were

also asked if they felt that the speaker's pronunciation was sufficient for him/her to TA an

undergraduate course (hereafter referred to as the 'TA question'). Finally, at the end of the

session, the raters identified a maximum of two speakers who stood out as the easiest and hardest

to understand overall.

Analysis of rater data

The qualitative data, which were elicited from open-ended questions in the rating

instrument, were analyzed using open coding, adapted from Strauss and Corbin (1998), with the

goal of generating categories that emerged from the data. The process was mostly cyclical rather

than linear, but essentially consisted of the following steps: (1) transcribing the data

chronologically by rater, with original spelling retained; (2) color-coding by speaker; (3)

grouping the color-coded data based on some perceived common thread (e.g., 'sound deletion');

(4) bolding meaning-laden words in the grouped utterances (e.g., 'no flow… thoughts…

disconnected'); (5) generating categories for the grouped data using the raters' language (e.g.,

'awkward pausing'); (6) subdividing categories (e.g., 'speech clarity' was divided into 'slurred

incomprehendably' and 'mumble/muffled'); and finally, (7) imposing the researcher's category

labels on the data (e.g., 'inability to immediately understand without effort'). A check for

consistency in categorization was carried out in accordance with procedures outlined in Johnson

and Christensen (2004) two weeks after the initial coding had been completed. Although the

order of comments within the categories was different, the basic categories generated were the

same, indicating a high level of intracoder reliability. All quantitative data were analyzed using

SPSS 13.0.

Results

Intelligibility

Table 1 shows the intelligibility ratings for the 18 undergraduate NSs, who are numbered in

the order of their mean ratings. Speaker 1, the highest rated speaker, scored 9% higher than

Speaker 2 and almost 20% above the group mean. Speaker 8, whose words were less than 50%

intelligible to the raters, scored over 20% lower than Speaker 7 and almost 30% below the group

mean. Although the ratings assigned to each speaker varied among raters (as can be seen in

standard deviations), the overall pattern of intelligibility ratings suggested that the speakers

represented varying degrees of speaking ability, at least with respect to the intelligibility measure

used here.

<PLEASE INSERT TABLE 1 ABOUT HERE>

Comprehensibility

Table 2 charts the speakers' comprehensibility scores. A Spearman rank-order correlation,

which was calculated to assess the level of consistency between intelligibility and

comprehensibility ratings, yielded a moderate positive relationship (r_s = .799). The only

difference in the order of ratings between Tables 1 and 2 is the reversal in position of Speakers 5

and 6.

<PLEASE INSERT TABLE 2 ABOUT HERE>

Figure 1 shows the frequency with which the raters identified the easiest and most difficult

speakers to understand at the end of the rating sessions. Evidence that Speakers 1 and 8

continued to hold the positions of the most and least comprehensible speakers respectively is

plain from their 'skyscraper' bar lines. Speaker 7 also appeared to the raters as being difficult to understand, albeit to a lesser extent than Speaker 8.

<div align="center"><PLEASE INSERT FIGURE 1 ABOUT HERE></div>

Researcher-coded intelligibility

The researcher's color-coded analysis of the speech data revealed that Speaker 1 was 100% intelligible (i.e., produced no unintelligible words). Conversely, Speaker 8 had considerably more unintelligible words than the other speakers, with approximately 1 in 6 words being color-coded 'unintelligible.' That still made her 82.46% intelligible in the intelligible words over total words ratio, however, far from the 46.94% that the raters assigned in their intelligibility ratings (Table 1). While the raters perceived a higher proportion of unintelligible words than the researcher had (the group means assigned by the raters and researcher were 76.59% and 96.80% respectively), there was, nonetheless, agreement on the negative effects of unintelligibility. Rater 6, for instance, who had assigned Speaker 8 an intelligibility score that was almost 15% above Speaker 8's mean score, made the following comment. (R6#S8 = Rater 6's comments about Speaker 8).

> R6#S8 Easy to make out most words (60%), but difficult to make sense overall since 60% is all you understand!

This statement that Speaker 8's words were easy to 'make out' but that the overall meaning was difficult to comprehend conforms with the researcher's speech sample analysis. The proportion of 1 in 6 words being unintelligible is debilitating in terms of allowing the listener 'to make sense overall,' even if the majority of words were discernable. Although the raters may have misjudged the proportion of unintelligible words that Speaker 8 had uttered in the same way that it is difficult to guess the number of green jellybeans in a jar, they were still able to identify the most and least intelligible speakers. Moreover, the order of the speakers in the raters'

intelligibility ratings closely coincided with the order derived from the researcher's color-coding, as attested by a strong Spearman rank-order correlation coefficient ($r_s = .905$). The differences were that the researcher found Speaker 6's speech to be more intelligible than Speaker 4's and 5's speech. The raters, however, viewed Speaker 5 to more intelligible than Speaker 6, although this order was reversed in their comprehensibility ratings. Admittedly, Speakers 4, 5, and 6, who placed in the middle of the group, were more difficult to distinguish than the more extreme cases.

There seems to be little consensus among the raters about Speaker 4's intelligibility or comprehensibility. Figure 1 shows that some raters cited Speaker 4 as the easiest to understand overall ($N = 3$), while others found him to be the most difficult ($N = 2$). Speaker 4, in fact, poses a challenge to the definition of intelligibility used in this study, since it does not make the distinction between important and unimportant words. In other words, there is no specification about which words must be intelligible – the criterion is just that the words must be immediately understood without recourse to guessing. This point is illustrated in Figure 2, which constituted the only instance of unintelligible speech in Speaker 2's passage.

<PLEASE INSERT FIGURE 2 ABOUT HERE>

The displaced stress on [dɪstɹɪbˈjutər] was color-coded for unusual pronunciation but was not found to be unintelligible. What was unintelligible was the rapidly spoken 'I'm gonna I'm gonna' [»ʌm ɴ ənæ ɴ ʌm ɴ ənæ«], which took many listenings to decipher. It is possible to see how a NS or NNS may have skimmed over these words, not enunciating them clearly. This seems to be a less severe instance of unintelligibility than an example of double substitution that arose when Speaker 8 said 'slipshot' instead of 'snapshot' – a word much more fundamental to the meaning of her utterance. The definition of intelligibility employed in this study, however, is impervious to such differences (see Zielinski, 2006).

Although all raters received the same 'treatment,' (instructions, teaching session, and

practice rating), they likely did not all interpret 'intelligibility' in the same way. The following

comments, grouped under the heading 'a feature of the speaker's pronunciation that, while

noticeable or irritating, does not affect overall intelligibility' shed light on the matter:

> R1#S2 other than mispronunciation of certain words, fairly easy to understand.

> R7#S5 Easy to understand except for pauses & mumbling

> R13#S1 He overpronounced words and spoke slowly but it didn't really hinder
> comprehension.

> R10#S1 understood all he said, but he spoke too slowly and had no intonation in his
> voice which made it quite obnoxious

> R18#S2 pronunciation is irritating but does not effect clarity of what is said.

The comments show that at least some raters were cognizant of the fact that some aspects

of a given speaker's pronunciation, or speech style, while perceptually salient, do not impede

intelligibility.

The TA question

Let us now turn to the TA question, which was positively correlated with the raters'

intelligibility ($r_s = .659$) and comprehensibility ratings ($r_s = .776$). Table 3 shows that no raters

felt that Speaker 8's pronunciation was sufficient to TA an undergraduate course, although 1 in 6

raters indicated that they were uncertain. Notably, Speaker 1's position at the most highly rated

speaker is usurped here by Speaker 2. Although seemingly negligible, this demotion can be

explained by examining individual raters' comments in conjunction with their quantitative

judgments. Although Rater 16 had marked that Speaker 1 was 100% intelligible and 'very easy'

to understand, he responded negatively to the TA question, elaborating 'way too slow to be a

TA.' This parallels comments from Raters 3 and 11, both of whom identified Speaker 1 as 100%

intelligible, 'easy,' and 'very easy' to understand respectively, and, in contrast to Rater 16, 'yes'

to the TA question.

R3#S1 Far too slow, as a TA, he would be quite boring.

R11#S1 only a few problematic words. but would be quite annoying as a TA.

<PLEASE INSERT TABLE 3 ABOUT HERE>

Rater 18, who also marked Speaker 1 as 100% intelligible and had even guessed that his L1 was English (apparently not detecting a foreign accent), neglected to fill out the subsequent comprehensibility and TA questions (missing data). In the comments, however, he wrote, 'sounds like someone with a good accent and no "orating" skill at all.'

Figure 3 shows that Speaker 1 had the slowest overall speaking rate on the *TSE* items that the raters assessed (1.82 syllables/second compared to the group mean of 2.44 syllables/second). His articulation rate, however, was the fastest of the group (4.84 syllables/second compared to the group mean of 3.40 syllables/second). That is, during his utterance time, he spoke quickly relative to the other speakers, approaching the optimal rate for NNSs that Munro and Derwing (2001) describe. However, frequent and long pauses (albeit at the end of thought groups) resulted in a high overall pause time. In the picture narrative, for instance, pausing accounted for 36.86 seconds of his 60.17 seconds of recorded speech. Perhaps this can, in part, account for the following comments:

R5#S1 I feel frustrated waiting for him to get on with what he is trying to say.

R11#S1 hard to link parts of the sentence together

R14#S1 Though slow, the choice of words is excellent and the effect is soothing.

<PLEASE INSERT FIGURE 3 ABOUT HERE>

Although there is more at play here than speech/articulation rates, these variables likely affected raters' responses to the TA question.[6] Notably, however, Speaker 1's rates may have been quite different on a task that had, for instance, required him to speak from prepared notes

on a familiar topic.

In sum, although some raters judged Speaker 1 to be 100% intelligible, they seemed reticent about his TAing an undergraduate course. However, the majority of raters responded positively to the TA question for him and for most other speakers. Six out of 8 speakers received more 'yes' responses than 'no,' although opinion for Speakers 5 and 6 was split between 'yes' on one hand, and 'no' and 'not sure' on the other.

Discussion

The data suggest that intelligibility, as defined for this study, is an adequate assessment criterion in the academic domain, although with minor qualifications. Speakers who performed well in intelligibility and comprehensibility ratings were also judged to have adequate pronunciation to TA an undergraduate course. Moreover, the relative order of intelligibility assigned by the raters after they had listened to 2 minutes of speech closely coincided with the researcher's opinion after multiple listenings in the analysis of the speech data. However, even for speakers found to be highly intelligible, certain factors about their speech made some raters wary of those speakers TAing undergraduate courses. In other words, there is evidence that intelligibility is a necessary but not a sufficient condition to TA an undergraduate course. This finding is supported by previous ITA research that has shown that there is more to teaching than accent and intelligibility alone (e.g., Tyler, 1992).

The results suggest that intelligibility, as defined for this study, is well-placed to be a useful assessment criterion for ITA screening, although the sample size precludes generalization of the results to other contexts. The word-level focus in the theoretical definition of intelligibility, adapted from the *EPT*, facilitated the construct's quantification in its operational definition, since the number of words that are understood is more countable than a definition that incorporates meaning. At the same time, the operational definition captured the raters' initial impressions of

the speech samples by getting them to plot points on a rating scale after just one listening.

Procedures in the analysis of the speech data, however, were less compatible with the theoretical

definition, since it was difficult to judge what was and was not immediately understandable after

the multiple listenings required for phonetic and suprasegmental transcriptions.

Two major arguments challenge the way that intelligibility was measured in this study. On

one hand, it was perhaps not objective enough to be reliable. The raters' intelligibility scores

might not have been as variable had a different measure (e.g., a dictation) been employed. On the

other hand, it could be argued that in the real world of communication, what is important is not

to understand every single word that is uttered, but rather the overall meaning or gist of the

message. This point throws into question the real-world value of a definition of intelligibility that

scrutinizes individual words, and would lean toward a concept of the term that embodies the

broader goal of communication (e.g., Morley, 1994).

The data suggest that there may be a threshold level of intelligibility and that Speaker 8 is

below that threshold, although a precise value for this was not defined. Perhaps one way of

determining an intelligibility threshold in a mixed methods study is to establish a cut-off score

based on the number of standard deviations that a speaker's score is from the mean (set at one or

two for a standard score, depending on the desired severity). The result could then be

triangulated with other sources of evidence (e.g., raters' comments). In this data set, for example,

the only speaker with a standard intelligibility score of one standard deviation or more below the

mean was Speaker 8 ($z = -2.03$).

One implication of the study is that it may be advantageous to assess the pronunciation of

NNS graduate students and prospective ITAs. While the *TSE* and *TOEFL* measure pronunciation

as one of several components of speaking, the development of an assessment instrument that has

a more sophisticated approach to evaluating pronunciation might be useful (Koren, 1995). Such

an instrument would need to establish a proficiency standard for the pronunciation of NNS

graduate students and/or prospective ITAs. Although not focusing on pronunciation specifically,

Saif (2002) reported on a needs-assessment approach to the construction of a general speaking

test at the University of Victoria. She employed Bachman and Palmer's (1996) test development

framework to foster the link between the purpose of the assessment, context, the domain of the

target language use, etc. This methodical approach to test development might be a useful starting

point at institutions that endeavour to construct their own pronunciation assessment instruments

for use at that institution.

Another issue worthy of consideration relates to the raters' willingness to engage in the

listening/rating tasks as receptive (as opposed to reluctant or hostile) interlocutors. To elaborate,

intelligibility presupposes the existence of both a speaker and a listener. Building on Smith and

Nelson's (1985) argument that intelligibility is interactional rather than speaker- or hearer-

based,[7] Rajadurai (2007) argues that intelligibility is co-constructed – a negotiated process

between the speaker and listener that is both dynamic and firmly embedded in its sociocultural

and interactive context (cf. Duranti's concept of the audience as co-author, 1986). In fact,

previous research has shown that intelligibility goes beyond the precision of the sounds that the

speaker produces to incorporate linguistic and non-linguistic factors related to listeners' attitudes

and expectations (e.g., Lindemann, 2002; Rubin, 1992). Lippi-Green argues that listeners'

negative attitudes towards accented speech can lead them to abdicate their share of the

communicative burden and even to claim that a NNS with good pronunciation and

communication skills is unintelligible (1994; 1997). Such listener attitudes have the potential not

only to adversely impact their interactions with NNSs, but also to bias their assessments of non-

native speech. The scoring rubrics of some assessment instruments also incorporate the notion of

listeners' willingness to engage in communicative interaction. The self-assessment grid for the

spoken interaction component of the Common European Framework, for example, invokes the concept of 'locuteur bienveillant,' where interaction is sustained 'provided the other person is prepared to repeat or reparhse things at a slower rate of speech and help me formulate what I'm trying to say' (Council of Europe, 1991, p. 26). This implies that the speakers' performance at the specified band level is contingent upon the interlocutor's good will, ability modify his/her speech, etc.

Ratings are, by nature, subjective, and the validity of the ratings in this study hinges, to some extent, on the raters' willingness to listen attentively to the speech and to co-construct meaning, even at the word level. Notably, the social context of the study is somewhat decontextualized and inauthentic. Although the NNSs were informed that the *TSE* is commonly used as an ITA screening instrument, their performance in this low-stakes research setting is different than it might have been had their prospective teaching jobs been at stake in a genuine ITA testing situation. Further, the construct definition of the tape-mediated *TSE* includes communication but not interaction (Educational Testing Service, 2001), and the negotiation of meaning is not built into the test (Caldwell & Samuel, 2001). Consequently, it is likely that the speakers in this study did not modify their output for their audience and were even unclear as to who their audience was. Thus, in both its real-world administration and in this study, the *TSE* tasks stray from the bilateral communication that ITAs are likely to encounter with their undergraduate students in real-life tasks (Saif, 2002), rendering the test's validity for assessing oral interaction questionable (Luoma, 2001). Furthermore, the fact that the stakes in this study were low may have influenced the raters' performance. Their attention to the speech and approach to the assessment would likely have been different in a real-world ITA context, where their ability to understand the ITA (and/or professor) is likely a factor in their success in the course.

Future research could explore the effect of altering the stakes of understanding the speaker's words by employing the matched-guise technique. For instance, four groups of undergraduate student raters could listen to and evaluate the recorded speech of an ITA delivering a lecture on a novel topic in their field of study. The first and second groups would be told that the individual in the recording is their future teaching assistant and that they will be responsible for the material covered in the lecture; the third and forth groups would be told that the individual is a lecturer at a neighboring university and that the material is for their interest only. The first and third groups would listen to a highly-accented version of the simulated lecture; the second and forth groups would listen to a moderately-accented version of the same lecture recorded by the same speaker. In addition to investigating group differences on intelligibility ratings, the researcher could examine differences in the quality of the comments of students from the different groups. This could shed light on the effect of manipulating the stakes of understanding and the degree of the speaker's accentedness on raters' approach to the assessment task, their perceptions of intelligibility, and their apparent willingness to attend to the speech.

Concluding remarks

In the notes that accompany her *Speech intelligibility/communicability index*, Morley suggests, 'try to listen to the speech sample as if you were an untrained language listener. Err on the conservative side with consideration of the "lay" listeners whom the student will meet' (1994, p. 77). Echoing Morley, Porter and Weir (1997) argue that 'at least part of the validation of criteria for assessing proficiency in pronunciation must be the gathering of information on what ordinary (non-linguist, non-applied-linguist, non-language-teacher) language users react to in the pronunciation of learners, and what the nature of the reaction is' (p. 25). Not only are undergraduate raters the lay listeners that Porter and Weir and Morley (1994) call for, but they

are also main stakeholders in the ITA context. An undergraduate student voice might be

instrumental in determining an acceptable pronunciation assessment standard for ITAs in the

screening process.

Bio statement

Talia Isaacs is a doctoral student in the Department of Integrated Studies in Education at McGill University. Her interest in pronunciation stems from her L2 teaching experience coupled with phonetics training in diction courses for opera singers.

Contact

Talia Isaacs

Department of Integrated Studies in Education

Faculty of Education, McGill University

3700 McTavish Street

Montreal, Quebec, Canada H3A 1Y2

Tel: 514-739-2868

E-mail: talia.isaacs@mcgill.ca

Acknowledgements

Notes

[1]   The author is indebted to Dr. Wayne Dickerson, developer of the Oral Interview section of the *EPT* at the University of Illinois at Urbana-Champaign, for providing information on *EPT* assessment procedures (personal communication, March 31, 2005).

[2]   All transcriptions and coding of the speech data were completed prior to the rating sessions. In other words, the raters' judgments had no bearing on the researcher's assessments.

[3]   The *TSE* tasks are reproduced in Isaacs (2006).

[4]   The word 'intelligibility' does not appear on the rating scheme since the pilot session revealed that the raters may not be familiar with the term and that the construct could be explained without direct reference to it. The word-level focus was deemed appropriate for this population of raters since words are a unit with which lay listeners are familiar (as compared to syllables or phonemes).

[5]   The assumption behind the use of 'comprehensibility' in this study was that it should correlate with 'intelligibility.'

[6]   Although Speakers 4 and 8 also have visibly higher articulation rates than the other speakers in Figure 3 (3.76 and 3.66 syllables/second respectively), the effect of their accelerated articulation rates were likely different than for Speaker 1. Half of the raters commented on Speaker 4's use of fillers, and the researcher described his speech using the analogy of 'a motor of a stalled car which sputters a bit, madly accelerates as if catching on only to sputter again a few seconds later' in a research journal. The raters described Speaker 8's 'irratic speach' (sic) 'awkward pauses,' and a jarring stop-and-start effect. Unlike Speaker 1, her pauses were abrupt and not at the end of thought groups.

[7]   This view is directly at odds with Fayer and Krasinski's contention that intelligibility is 'hearer-based' (1987, p. 313).

References

Abercrombie, D. (1949). Teaching pronunciation. *English Language Teaching*, *3*, 113-122.

Anderson-Hsieh, J., Johnson, R. & Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning*, *42*, 529-555.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Caldwell, K., & Samuel, C. (2001). A Review of the Test of Spoken English (TSE). *Canadian Modern Language Review*, *58*, 319-326.

Catford, J. C. (1950). Intelligibility. *English Language Teaching, 5*, 7-15.

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

Cheng, L., Myles, J., & Curtis, A. (2004). Targeting language support for non-native English-speaking graduate students at a Canadian university. *TESL Canada Journal, 21*, 50-71.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge Cambridge University Press.

Dalton, C. & Seidlhofer, B. (1994). *Pronunciation*. Oxford: Oxford University Press.

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, *20*, 1-16.

Duranti, A. (1986). The audience as co-author: An introduction. *Text, 6*, 239-247.

Educational Testing Service. (1995). *Test of Spoken English: Standard-setting manual*. Princeton, NJ: Educational Testing Service.

Educational Testing Service. (2001). *TSE and SPEAK score user guide*. Retrieved February 23,

2008, from http://ftp.ets.org/pub/toefl/008659.pdf

Fayer, J. M. & Krasinski, E. (1987). Native and nonnative judgments of intelligibility and

irritation. *Language Learning*, *37*, 313-326.

Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly*, *39*,

399-423.

Gass, S., & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of

nonnative speech. *Language Learning*, *34*, 65-89.

Gimson, A. C. (1980). *An introduction to the pronunciation of English* (3rd ed.). London:

Edward Arnold.

Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of

suprasegmentals. *TESOL Quarterly*, *38*, 201-223.

Hoekje, B., & Williams, J. (1992). Communicative competence and the dilemma of international

teaching assistant education. *TESOL Quarterly, 26*, 243-269.

Isaacs, T. (2006). *Towards defining a valid assessment criterion of pronunciation proficiency in*

non-native English speaking graduate students. Unpublished Master's Thesis, McGill

University, Montreal, Canada.

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

University Press.

Jenkins, J. (2002). A sociolinguistically based, empirically researched pronunciation syllabus for

English as an International Language. *Applied Linguistics, 23*, 83-103.

Johnson, B., & Christensen, L. B. (2004). *Educational research: Quantitative, qualitative and*

mixed approaches (2nd ed.). Boston: Pearson Education.

Kenworthy, J. (1987). *Teaching English pronunciation*. London: Longman.

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

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*, 419-441.

Lippi-Green, R. (1994). Accent, standard language ideology, and discriminatory pretext in the courts. *Language in Society, 23*, 163-198.

Lippi-Green, R. (1997). *English with an accent: Language, ideology and discrimination in the United States*. London: Routledge.

Ludwig, J. (1982). Native speaker judgments of second language learners' efforts at communication: A review. *Modern Language Journal, 66*, 274-283.

Luoma, S. (2001). A Review of the Test of Spoken English (TSE). *Language Testing, 18*, 225-234.

Morley, J. (1994). A multidimensional curriculum design for speech-pronunciation instruction. In J. Morley (Ed.), *Pronunciation pedagogy and theory* (pp. 64-91). Alexandria, VA: Teachers of English to Speakers of Other Languages.

Munro, M. J., & Derwing, T. M. (1998). The effects of speaking rate on listener evaluations of native and foreign accented speech. *Language Learning, 48*, 159-182.

Munro, M. J., & Derwing, T. M. (1999). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *49*, 285-310.

Munro, M. J., & Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech: The role of speaking rate. *Studies in Second Language Acquisition, 23*, 451-468.

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

Pickering, L. (2001). The role of tone choice in improving ITA communication in the classroom. *TESOL Quarterly*, *35*, 233-255.

Porter, D. & Weir, C. (1997). Valid criteria for the assessment of proficiency in pronunciation. *Speak Out*, *20*, 22-28

Rajadurai, J. (2007). Intelligibility studies: A consideration of empirical and ideological issues. *26*, 87-98.

Rubin, D. L. (1992). Nonlanguage factors affecting undergraduates' judgments of non-native English-speaking teaching assistants. *Research in Higher Education, 33*, 511-531.

Saif, S. (2002). A needs-based approach to the evaluation of the spoken language ability of international teaching assistants. *Canadian Journal of Applied Linguistics*, *5*, 145-167.

Saif, S. (2006). Aiming for positive washback: A case study of international teaching assistants. *Language Testing*, *23*, 1-34.

Smith, L. E. (1992). Spread of English and issues of intelligibility. In B. B. Kachru (Ed.), *The other tongue: English across cultures* (2nd ed., pp. 75-90). Urbana, IL: University of Illinois Press.

Smith, L. E., & Nelson, C. I. (1985). International intelligibility of English: Directions and resources. *World Englishes, 4*, 333-342.

Strauss, A., & Corbin, J. (1998). *Basics of qualitative research* (2nd ed.). Thousand Oaks, CA: Sage.

Turner, C. E., & Upshur, J. A. (2002). Rating scales derived from student samples: Effects of the scale maker and the student sample on scale content and student scores. *TESOL Quarterly*, *36*, 49-70.

Tyler, A. (1992). Discourse structure and the perception of incoherence in international teaching assistants' spoken discourse. *TESOL Quarterly*, *26*, 713-729.

Zielinski, B. (2006). The intelligibility cocktail: An interaction between speaker and listener ingredients. *Prospect, 3*, 20-43.

Table 1

Mean intelligibility ratings

Speaker	Mean (%)	SD	Range (%)
1	95.22	8.54	75 - 100
2	86.28	13.72	50 - 100
3	85.78	12.62	50 - 100
4	79.06	18.18	25 - 100
5	76.56	13.97	50 - 98
6	75.17	15.09	50 - 100
7	67.67	18.24	30 - 88
8	46.94	18.51	8 - 86
Group	76.59	20.19	8 - 100

Table 2
Frequencies of comprehensibility scores

Speaker	Very Difficult	Difficult	Easy	Very Easy	Missing
1	-	1	8	8	1
2	-	2	15	1	-
3	-	4	8	4	2
4	1	4	9	4	-
6	-	7	7	1	3
5	1	10	6	1	-
7	1	11	6	-	-
8	9	9	-	-	-
Total	12	48	59	19	6

Note. Means are not displayed due to missing data.

Table 3
Is the Speaker's Pronunciation Adequate to TA.?

Speaker	Yes (%)	No (%)	Unsure (%)	Missing
2	77.8	5.6	16.7	-
1	76.5	11.8	11.8	1
3	61.1	5.6	33.3	-
4	61.1	16.7	22.2	-
6	50.0	16.7	33.3	-
5	50.0	27.8	22.2	-
7	27.8	38.9	33.3	-
8	-	83.3	16.7	-
Group	50.0	25.7	23.6	1

Figure 1
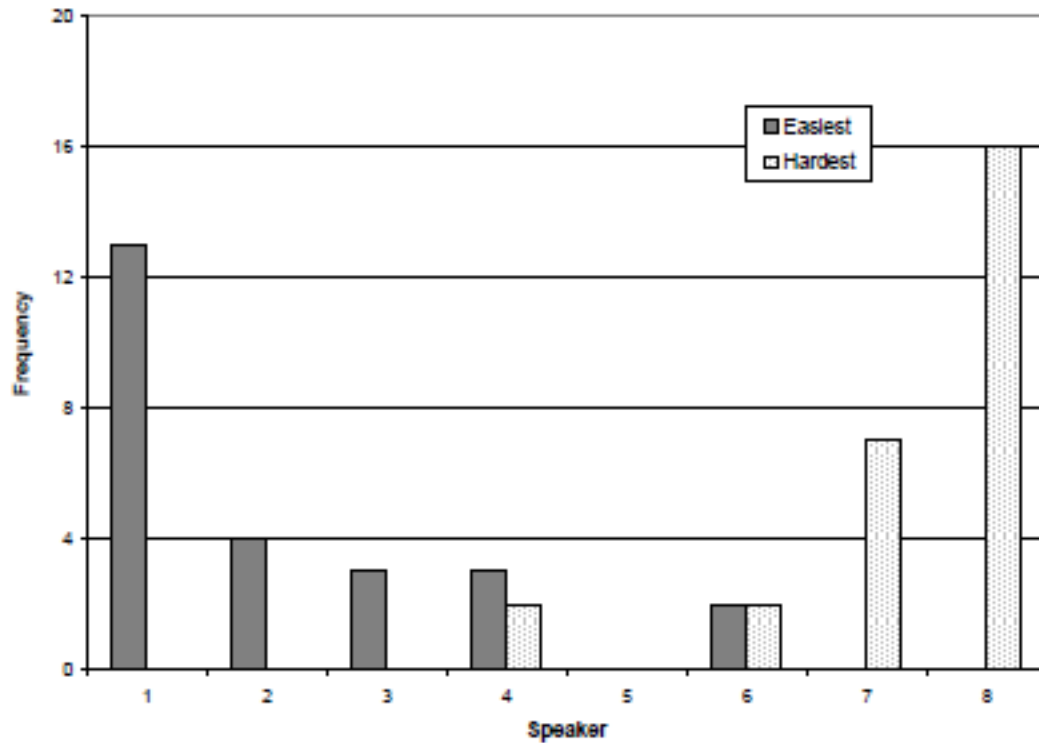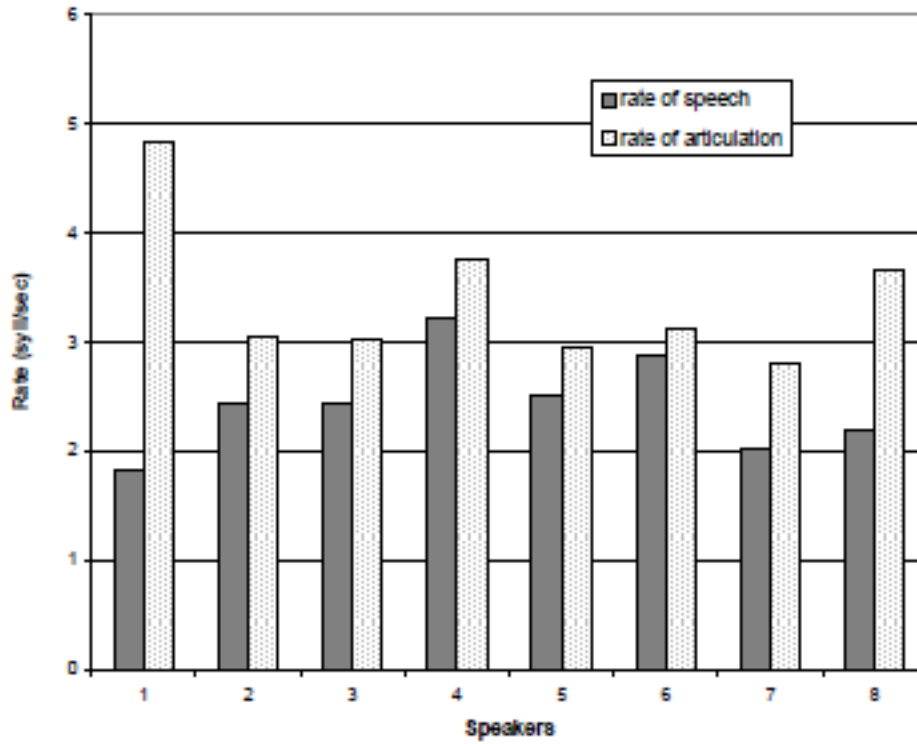*Easiest and hardest speakers to understand overall*

Figure 2
Identifying unintelligible speech

[aj_dɪstɹɪb'ʲᵘtər_ə_æ_æ-d _ 'bifɔɹ bʌʔ aj_'nidəd_ty tʃejndʒ sʌm
  i distributed a ad before but i needed to change some

'sʌmθɪŋ (1.) sow_ »ʌm_ənæ_ʌm_ənæ« tɛl_jy_wʌɾ_aj tʃ*ejnʤd* _]
something so um onna um onna tell you what I changed…

Figure 3
*Speech and articulation rates for rated tasks*

Appendix 1

First section of the rating instrument

In this section, you will be asked to rate the pronunciation of a non-native English speaker. As you listen to this person's speech for the first time, see if you can understand every single word that he/she says.

Remember: By "understand" I mean that you are able to comprehend each word immediately so you do not have to guess at words.

On the scale below, mark approximately what percent of the speaker's words you are able to understand with an "X."

| 0% | 25% | 50% | 75% | 100% |
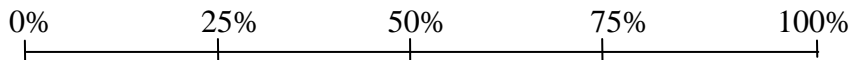
As you listen again, try to identify whether any of the below features hindered your ability to understand the speaker's words.

☐**a. <u>speech clarity</u>** - the speaker:
  ☐ overpronounces words (articulates each syllable painstakingly)
  ☐ mumbles/ eats words (speech is unclear or muffled)

☐**b. <u>rate of speech</u> -** the speaker:
  ☐ speaks too fast
  ☐ speaks too slowly

☐**c. <u>pitch</u>** – the speaker's pitch:
  ☐ changes too often from high to low
  ☐ doesn't change enough/ is too monotone

☐**d. <u>sentence rhythm</u>** - the speaker:
  ☐ fails to distinguish between important and unimportant words in the sentence
  ☐ fails to link sounds between words (e.g., doesn't connect the "z" sound to the "a" in applez**N**N and oranges")

☐**e. <u>word stress</u>** – the speaker:
  ☐ often doesn't get the syll-A-ble right
  ☐ often doesn't distinguish between strong and weak syllables

☐**f. <u>individual consonant/ vowel sounds</u>** – the speaker:
  ☐ substitutes problematic sounds for ones that are easier to pronounce (e.g., says "sink" instead of "think" or "heat" instead of "hit")
  ☐ adds sounds or deletes sounds (e.g., says "sundly" instead of "suddenly" or "warem" instead of "warm")

*☐NONE

(1) Now, rank order the top 3 features that hindered your ability to understand the speaker's words by placing a number in the big box, where 1 is for the most hindering feature.
(2) For the 3 features you rank ordered, check the most prominent problem (i.e., the one that stands out to you the most) in the small box below the letter. You may only check one option.
(3) If none of features #a-f interfered with your ability to understand the speaker's words, leave those boxes blank and mark an "X" in the NONE *box at the bottom of the page.*

Appendix 2

Textual symbols for stress, rhythm, pitch, and intonation

' primary stress

underlined multi-syllabic word where there is no clear main stress

• words emphasized in the sentence

◇ lack of distinction between important and unimportant words

subscripted low pitch

superscripted high pitch

subscripted indented H→L tone sliding

superscripted indented L→H high rise intonation

Y   Z monotone, no inflection

↑+3 ↓-6 pitch rises ↑ or falls ↓ by the musical pitch internal indicated

Textual symbols for pauses and tempo

»speeds up«

«slows down»

_ brief unmeasured pause (<1 second)

(1.2.3) pause length to the nearest second when duration is ≥1 second

- consonant/ vowel sound prolonged at same pitch

/ last two sounds detached (no linking)

! staccato (abrupt detached articulation)

 N legato and/ or linking from one sound to another

Textual symbols for segmentals (non-IPA)

sound deletion

Other textual symbols

((coughs))

Color-coding for intelligibility

green – unusual pronunciation that does not affect intelligibility

red – unintelligible pronunciation