

IOWA STATE UNIVERSITY

Digital Repository

English Publications

English

2014

From One to Multiple Accents on a Test of L2 Listening Comprehension


Gary Ockey

Iowa State University, gockey@iastate.edu

Robert French

Educational Testing Service

Follow this and additional works at: http://lib.dr.iastate.edu/engl_pubs

 Part of the [Bilingual, Multilingual, and Multicultural Education Commons](#), [Curriculum and Instruction Commons](#), [Educational Assessment, Evaluation, and Research Commons](#), and the [Educational Methods Commons](#)

The complete bibliographic information for this item can be found at http://lib.dr.iastate.edu/engl_pubs/83. For information on how to cite this item, please visit <http://lib.dr.iastate.edu/howtocite.html>.

This Article is brought to you for free and open access by the English at Iowa State University Digital Repository. It has been accepted for inclusion in English Publications by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.

From One to Multiple Accents on a Test of L2 Listening Comprehension

Abstract

Concerns about the need for assessing multidialectal listening skills for global contexts are becoming increasingly prevalent. However, the inclusion of multiple accents on listening assessments may threaten test fairness because it is not practical to include every accent that may be encountered in the language use domain on these tests. Given this dilemma, this study aimed to determine the extent to which accent strength and familiarity affect comprehension and to provide a defensible direction for assessing multidialectal listening comprehension. A strength of accent scale was developed, and one US, four Australian, and four British English speakers of English were selected based on a judgment of their strength of accent. Next, TOEFL test takers ($N = 21,726$) were randomly assigned to listen to a common lecture given by one of the nine selected speakers, and respond to six comprehension items and a survey designed to assess their familiarity with various accents. The results suggest that strength of accent and familiarity do affect listening comprehension, and these factors affect comprehension even with quite light accents.

Disciplines

Bilingual, Multilingual, and Multicultural Education | Curriculum and Instruction | Educational Assessment, Evaluation, and Research | Educational Methods

Comments

This is a pre-copyedited, author-produced PDF of an article accepted for publication in *Applied Linguistics* following peer review. The version of record "From One to Multiple Accents on a Test of L2 Listening Comprehension," (2014) 1-24 is available online at: <http://dx.doi.org/10.1093/applin/amu060>.

From One to Multiple Accents on a Test of L2 Listening Comprehension

Gary J. Ockey

Iowa State University

Robert French

Educational Testing Service

Abstract

Concerns about the need for assessing multidialectal listening skills for global contexts are becoming increasingly prevalent. However, the inclusion of multiple accents on listening assessments may threaten test fairness because it is not practical to include every accent that may be encountered in the language use domain on these tests. Given this dilemma, this study aimed to determine the extent to which accent strength and familiarity affect comprehension and provide a defensible direction for assessing multidialectal listening comprehension. A strength of accent scale was developed, and one US, four Australian, and four British English speakers of English were selected based on a judgment of their strength of accent. Next, TOEFL test takers ($N = 21,726$) were randomly assigned to listen to a common lecture given by one of the nine selected speakers, and respond to six comprehension items and a survey designed to assess their familiarity with various accents. The results suggest that strength of accent and familiarity do affect listening comprehension, and these factors affect comprehension even with quite light accents.

From One to Multiple Accents on a Test of L2 Listening Comprehension

Concerns about the use of only one select English accent for assessing second language (L2) listening comprehension are becoming increasingly prevalent (Harding 2011; Taylor and Geranpayeh 2011; Abeywickrama 2013). Critics of this practice of using only one select variety of English point to the changing demographics in many English speaking contexts and argue that L2 listening assessments need to reflect these changes. These voices contend that given the variety of accents which may be encountered in these contexts, it may be necessary to have multidialectal listening skills to communicate successfully in English speaking contexts. In North American universities, for instance, it is not uncommon for 20-30% of teaching assistants to have accents different from a standard United States variety (Department of Institutional Research, 2007). It follows that scores based on listening assessments which measure comprehension of only one selected accent may not reflect how well these test takers can function in such a multidialectal language use domain. To ensure that test takers are prepared for such diverse contexts, an argument can be made for including multiple accents on listening assessments designed to determine the extent to which test takers will be able to communicate in such environments.

Others express concerns about the use of a variety of accents on L2 listening comprehension assessments because some test takers may be unfairly disadvantaged (Elder and Davies 2006; Elder and Harding 2008; Taylor and Geranpayeh 2011). These voices point out that it would not be practical for each form of an assessment to include every type of accent that could be encountered in the target language use domain, which refers to the “situation or context in which the test taker will be using the language outside of the test itself,” (Bachman and Palmer, 1996, p. 18) and sampling from a large number of accents could be unfair because those

unfamiliar with the accent selected for any given test form may be disadvantaged (Field 2004; Taylor 2006). Thus, while assessing listening comprehension with speakers who have homogeneous accents may underrepresent the listening construct, including speakers with multiple accents may result in unfairly disadvantaging some test takers. Given this dilemma, the aim of this study was to determine the extent to which strength and familiarity of accent affect comprehension and provide a defensible direction for assessing multidialectal listening comprehension for listening assessments.

Literature review

This section begins with a discussion of what ‘accent’ is and how it has been defined. Research findings on the effects of familiar and unfamiliar accents on assessment scores are then discussed.

A construct definition of accent

Although a number of researchers have proposed definitions for accent, agreement on a definition has not yet been reached (Pennington 1996; Derwing and Munro 2009). This lack of agreement reflects the various purposes for which the term accent has been used. Some definitions aim to delineate the various linguistic features that the term accent might suggest. Harding (2011) indicated that segmental and suprasegmental differences in pronunciation, including variation in vowels and consonant sounds at the segmental level, and stress and intonation at the suprasegmental level, combine to shape an accent.

Other researchers have used sociolinguistic definitions which rely on listener judgments to define accent. Derwing and Munro (2009) define English speakers’ accents in relation to the local variety, that is, ‘the ways in which their speech differs from that local variety of English and the impact of that difference on speakers and listeners’ (476). In other words, for Derwing

and Munro, accent is defined in terms of how a speaker's spoken language sounds to others and the way in which a speaker's sounds affect the listeners and the speaker, compared to those who are identified as users of the local speech variety. Derwing and Munro (2009) discuss accent in terms of its salience, or how different it is perceived to be from the local dialect, and comprehensibility, that is 'the listener's perception of how easy or difficult it is to understand a given speech sample' (478). Their purpose in making this distinction was to emphasize that speakers who sound different from the local dialect are not necessarily less comprehensible than those speakers who have speech patterns that are judged to be the same as the local dialect. Prior research provided empirical evidence that these two dimensions are related, but to some extent distinct constructs (Munro and Derwing 1995a; Derwing and Munro 1997). Given these findings, it is crucial when defining accent to consider both the perceived difference from the local dialect and the extent to which listeners judge the speech to be comprehensible.

In this study, accent is defined based on the definitions of other researchers, previous research findings, and the need for a definition in L2 assessment that emphasizes the relationship between the differences in speech patterns and perceptions of the extent to which these differences impact a listener's comprehension. Accent is also defined as the degree to which an individual's speech patterns are perceived to be different from the local variety, and how much this difference is perceived to impact comprehension of listeners who are familiar with the local variety. Therefore, the strength of an accent indicates the degree to which it is judged to be different than the local variety, and how it is perceived to impact the comprehension of users of the local variety.

Effects of accent on listening comprehension

Some studies have investigated the effects of ‘accent’ on listening comprehension and failed to find an effect. Abeywickrama (2013) had Brazilian, Korean, and Sri Lankan English learners take a multiple-choice (MC) listening test in which the input was delivered by a Chinese, Korean, Sri Lankan, or US speaker. She used a one-way between groups ANOVA with speaker’s country of origin as the independent variable and score on a multiple-choice speaking test as the dependent variable. She did not find a significant relationship between scores on the MC listening assessment and the speakers who delivered the inputs. Abeywickrama assumed that the speakers had an accent that reflected their country of origin, but did not provide a measure of accent in the study. Because there was no measure of strength of accent, or other related construct such as intelligibility, used in the study, it is not clear to what degree these accents were different from the local variety with which the test takers were familiar.

The majority of research, however, suggests that a speaker’s accent can affect listening comprehension scores (e.g., Eisenstein and Berkowitz 1981; Ekong 1982; Smith and Bisazza 1982; Anderson-Hsieh and Kohler 1988; Bilbow 1989). Of particular importance to the current study was the research conducted by Anderson-Hsieh and Kohler (1988) that considered strength of accent in the study’s design and compared the listening comprehension of 224 North American university students across four English speakers. The first language of three of the speakers was Chinese, and the fourth speaker was North American. The three first language (L1) Chinese speakers of English were judged to have different levels of speaking ability based on scores on the Test of Spoken English (180, 200, and 260), and a judgment of the speakers’ pronunciation. The results indicated that the university students had significantly higher comprehension scores for input delivered by the North American speaker than the Chinese speakers, and that the comprehension was lowest for the Chinese speaker who had the weakest

English proficiency and poorest pronunciation. Importantly, test takers comprehended the North American speaker significantly better than all of the L1 Chinese speakers. The Chinese speakers were assumed to have different accent strengths based on their country of origin and their English language proficiency derived from judgments of their oral proficiency—including pronunciation.

Studies like Anderson-Hsieh and Kohler's (1988) have provided evidence that 'accent' can affect listening comprehension, but that it is not necessarily the case that it does. A likely reason for these mixed findings is that, as Anderson-Hsieh and Kohler (1988) conclude, there are various types and strengths of accents, based on Derwing and Munro's (2009) definition, some of which affect the listening comprehension of some listeners and others which do not. These studies point to the need for research that provides a clearly defined and defensible measure of accent accompanied by an indication of the strength of accent that impacts listening comprehension.

Effects of accent familiarity on listening comprehension

Possibly the biggest threat to the validity of listening scores yielded from speakers with different accents is that some test takers may be advantaged by familiarity with a particular accent encountered on a test, while others may be disadvantaged because they are assigned to take assessments with accents with which they are not familiar. Bradlow and Brent (2008) suggested that such an effect could result from phonetic characteristics of speech which are known to be rather consistent across speakers who have the same accent.

A substantial body of research suggests that familiarity with an accent positively relates to comprehensibility (Gass and Varonis 1984; Derwing and Munro 1997; Adank, et al. 2009; Adank and Janse 2010). Adank, et al. (2009) found that listening to an unfamiliar English accent

resulted in lower scores for native-speakers of English than listening to the same input from an English speaker with a familiar accent. Two British English accents (Southern Standard and Glaswegian) and two groups of British listeners were included in the study. Both groups of listeners were familiar with the Southern Standard variety but only one was familiar with the Glaswegian variety. The results indicated that familiarity with the accent led to better comprehension. A follow-up study reported in the same paper, however, did not indicate that comprehension was debilitated by an unfamiliar accent but did provide evidence that processing time for comprehension was significantly greater for unfamiliar than familiar accents. The researchers conjecture that the conflicting results of the two studies were due to the fact that the ‘accents for the two speakers were less prominent’ for the unfamiliar accents in the second study. These conflicting findings underscore the importance of clearly defining and measuring accent to determine its effect.

Major et al., (2002) found a familiarity effect for Spanish speakers, whose comprehension was higher when listening to an English speaker with a Spanish accent compared to an English speaker with a Chinese accent. This shared L1 hypothesis did not hold for Chinese listeners, however, who did not better comprehend a Chinese than Spanish speaker. The speakers were assumed to have Spanish and Chinese accents based on their country of origin, and as the researchers pointed out, their findings may be mitigated by their failure to account for item difficulty and use of a defensible measure of strength of accent.

The shared L1 hypothesis effect has been investigated by other researchers, who have concluded that it is not necessarily the shared L1 that relates to higher comprehension. Rather, it is familiarity based on previous exposure to the accent that dictates comprehension (Smith and Bisazza 1982; Ortmeier and Boyle 1985; Tauroza and Luk 1997). Harding’s (2011) study took

this body of research a step further by investigating possible factors that impact comprehension of unfamiliar accents. After corroborating previous research with a finding for greater comprehension of familiar than unfamiliar accents, Harding suggests that this may be due to misperception and the lack of ability to distinguish phonetic information, or challenges with processing speech.

The research to-date has mostly been founded on the assumption that speakers from countries different than those from the local dialect have accents. Moreover, although various definitions and measures of accent have been used, the findings generally affirm that familiarity with an accent is an advantage, but that this is not necessarily the case for all contexts. It is likely that the mixed results stem from the ways in which accent has been defined, and an effect that occurs only beyond a particular threshold of familiarity, strength of accent, or a combination of both. Identification of such a threshold would provide insights to test developers who desire to make an English listening construct more multidialectal without unfairly impacting some test takers. To shed light on these issues, this study aims to answer the following research questions when defining accent for a second language assessment context:

RQ₁: What is the relationship between an L2 listener's comprehension and the strength of a speaker's accent?

RQ₂: What is the relationship between an L2 listener's comprehension and familiarity with an accent identified as the speaker's?

Method

Development of a Strength of Accent Scale

A measure was developed to operationally define the construct of accent strength, although some might contend that along with accent strength, it also measured effect of accent. The Strength of

Accent Scale, which was developed for this study, is based on salience and comprehensibility (Derwing and Munro 2009) as well as an understanding gleaned from Adank, et al's (2009) study that unfamiliar accents may lead to additional processing time even though comprehension is not decreased. Given that previous research indicated that listener judgments have been shown to be a very reliable approach to assessing accent and comprehensibility (Derwing and Munro 2009), the measure was developed based on listener judgments. The study was conducted as part of a larger project which aimed to determine the effects of expanding the listening construct from 'Standard United States English' to a multidialectal one. Thus, it was determined that 'Standard United States English' would be identified as the local variety to which other varieties of English would be compared.

The development of the Strength of Accent Scale went through a number of phases. After a draft scale was created based on salience, comprehensibility, and additional processing time, three small focus groups, which included both L1 English speakers and highly proficient L2 English speakers, listened to various speakers who participated in the study, and attempted to use the scale to judge the strength of their accents. Based on feedback from the focus groups, the Strength of Accent Scale was revised. The same focus groups then used the revised scale to rate the accents of a different set of speakers who participated in the study and provided further feedback, which resulted in another revision of the measure. The final version of the Strength of Accent Scale is provided in Appendix A.

Selecting speakers for the study

Twenty adult speakers auditioned to participate in this study which included nine British, nine Australian, and two speakers from the United States. In selecting these speakers, the researchers, who were familiar with the local United States speech variety, believed that their

speech varieties were noticeably different than a United States variety, but did not consider the speakers to have strong accents.

The speakers who auditioned to be in the study were given time to familiarize themselves with the script, and were provided with guidance during the recording sessions by members of the recording team responsible for directing the recording of TOEFL listening comprehension section inputs. After this training, the speakers were recorded reading one of two academic scripts that were approximately five minutes in length. These recordings were made using a state-of-the-art sound system designed to record listening comprehension stimuli used for high stakes assessments.

Two twenty-second clips from the recordings were then created for each speaker. The decision to use twenty-seconds of input was based on research that has shown unfamiliar speech varieties to require increased processing time (Munro and Derwing 1995b; Schmid and Yeni-Komshian 1999; Adank et al. 2009) and feedback from the focus groups who felt this was a reasonable amount of time to make a judgment about the strength of a speaker's accent. Based on feedback from the focus groups, a number of principles were followed in making these clips:

- 1) No two clips were the same. It was determined that after a listener knew what a speaker would say, it would not be reasonable to expect the same type of judgment when compared to the first time a clip was encountered.
- 2) Care was taken to avoid using clips from lectures that included low frequency vocabulary. Listeners might not know if they could not understand a word because of the accent or because they did not know the word's meaning.
- 3) Each clip began at the start of a sentence.
- 4) Segments that made little sense without additional context were avoided.

The 40 twenty-second clips (two for each of the 20 speakers) were then spliced into two

different random orders with a narrator providing directions on how to judge each speaker's accent based on the Strength of Accent Scale (see Appendix A).

One-hundred students and instructors were then asked to listen to the audio clips and use the Strength of Accent Scale (see Appendix A) to judge the accents of the potential speakers. These students and instructors were from one of three United States institutions, a large university on the west coast ($n = 33$), a small Midwestern community college ($n = 33$), and a large central US university ($n = 34$). The participants were diverse in terms of major area of study (business [9], humanities [30], natural sciences [30], social sciences [31]); gender (female [67], male [33]); status (graduate student [27], instructor [8], undergraduate student [65]); and first or second language English speaker (first [61], second [39]). All second language English speakers had advanced proficiency, based on meeting the language requirement for studying in mainstream English content courses. At each institution, judges were randomly assigned to two groups, and each group listened to a different order of the speech samples. Accent ratings, based on the average rating of the 100 students and instructors, for the speakers who auditioned to participate in the study are shown in Table 1. Speakers have been ordered based on strength of accent—from weakest to strongest.

Table 1

Accent rating of speakers who auditioned to participate in study

Gender	Country of Origin	Rating Frequency					Rating			
		1	2	3	4	5	First	Second	Mean	SD
Female	US	185	13	2	0	0	1.1	1.1	1.1	0.3
Male	US	179	20	1	0	0	1.1	1.1	1.1	0.3
Male	AUS	81	102	16	0	0	1.8	1.6	1.7	0.6
Male	AUS	69	117	13	0	0	1.6	1.8	1.7	0.6
Male	UK	63	121	15	0	0	1.7	1.8	1.8	0.6
Male	AUS	56	130	13	0	0	1.8	1.8	1.8	0.6

Male	UK	59	121	18	1	0	1.9	1.7	1.8	0.6
Male	AUS	52	125	20	2	0	1.8	1.9	1.9	0.6
Female	UK	49	125	21	3	0	1.9	1.9	1.9	0.6
Female	UK	39	134	24	2	0	1.9	2.0	1.9	0.6
Male	UK	40	123	31	4	0	2.0	2.0	2.0	0.7
Male	AUS	28	139	28	4	0	1.9	2.2	2.0	0.6
Female	AUS	30	128	37	2	1	2.0	2.2	2.1	0.7
Female	AUS	36	110	49	4	0	2.0	2.2	2.1	0.7
Female	AUS	23	130	37	8	0	2.1	2.2	2.2	0.7
Female	UK	21	129	42	6	0	2.2	2.1	2.2	0.6
Male	UK	22	115	54	6	1	2.2	2.3	2.2	0.7
Female	UK	26	70	87	15	1	2.5	2.5	2.5	0.8
Female	UK	9	96	69	21	3	2.9	2.2	2.6	0.8
Female	AUS	10	73	94	20	1	2.7	2.6	2.7	0.8
Total		1077	2121	671	98	7	2.0	2.0	2.0	0.6

Note. Bold indicates selected to participate in study.

As shown in Table 1, the two average ratings for each speaker were fairly consistent with the exception of the female British speaker, who is shown second from the bottom. Eighteen of the two ratings for each speaker were within .2 points, and one rating was within .3 points. The difference between the two ratings of the British female who had an accent strength of 2.6 was .7 points. Estimation of reliability of the accent measure based on a correlational approach suggested fair internal consistency of scores with a value of $\alpha = .69$. This rather marginal estimate of reliability likely reflects the fact that only speakers with a rather narrow range of accents were included in the speaker audition. Given the high consistency of the ratings from 19 of the 20 speakers, and the much lower consistency of the ratings of one speaker, it is likely that one of the audio clips was not judged accurately, but a review of the two clips did not reveal an obvious reason for the rating differences. The standard deviation of scores increases slightly as the strength of accent increases, possibly suggesting that listeners provided less similar ratings of stronger accents.

From these results, nine of the twenty speakers who auditioned to participate in the study were selected. Speaker selection was largely driven by the desire to include a range of accent strengths, and to have an equal distribution of males and females, and Australian and British speakers. One United States speaker, who was judged to have an accent representative of the local variety, was also selected.

Participants

The study included 21,726 TOEFL iBT test takers from 148 countries. All test takers who took TOEFL on two consecutive weekends were included in the study, suggesting that the sample was quite representative of the TOEFL iBT test taker population (see http://www.ets.org/s/toefl/pdf/94227_unlweb.pdf for a details about the TOEFL test taker population).

The speakers who were selected for the study are shown in bold font in Table 1. The British and Australian speakers had accent strengths which ranged from 1.7, a little more than half way between *not noticeable* and *noticeable* and 2.7, a little more than half way between *noticeable* and *required concentrated listening* (see Appendix A). Therefore, based on an overall average of ratings for each speaker, the 100 students and instructors, who judged the strength of accent of the speakers and resided in the United States, believed that none of the speakers in the study had accents that limited comprehensibility.

Materials

A monologic lecture, 686 words in length, was the stimulus to be comprehended by the test takers. The lecture was on a natural sciences topic, and described and considered several hypotheses about why sea gulls engage in a behavior known as ‘drop-catch behavior’. As test takers listened to the stimulus, several context photographs of the speaker appeared on the

computer screen. Additionally, written on a single blackboard were the names of two researchers whose study was being described in the lecture. Test takers were able to take notes as they listened to the lecture, and to use their notes when they answered the questions. The lecture was followed by six questions: a general idea question asking about the topic of the lecture, two detail questions asking about important points made in the lecture, a pragmatic understanding item asking about an opinion the professor expressed, and two connecting information items asking about the relationship between two pieces of information. Several criteria guided the selection of the particular lecture to be used. First, only lectures that had been developed and pretested for use on a TOEFL listening test were considered. Second, the lecture needed to be, at most, of average difficulty, compared to TOEFL listening sets, and have accessible content. Third, it should contain few, if any, technical terms, and few, if any, lexical items that would clearly be inappropriate for a British or Australian speaker's speech variety (for more information about the TOEFL iBT listening section see: <http://www.ets.org/toefl/ibt/about/content>).

The TOEFL listening test is presented to test takers in blocks of three sets of items. A single block consists of a conversation between two speakers, a monologic lecture, and an interactive lecture. There are five items based on the conversation and six questions for each lecture. Thus, for a single block of three sets of items, a test taker must answer seventeen questions. All speakers of the inputs of these sets of items spoke with the local United States variety of English. Scores on two complete sets of items ($k = 34$) with speakers who spoke the local United States variety of English were used as a covariate in the study to assure that test takers in each treatment group had equivalent listening abilities. The use of these 34 items in the study is further discussed in the *Testing Procedures* section.

Test takers were asked to respond to a questionnaire designed to assess their familiarity with the accents used in the study. The questionnaire included four questions about test takers' experience with United States, Australian, and British accents: experience with the accent in general, in face-to-face communication, through the media, and with a teacher (see Appendix B). The questionnaire also included items about familiarity with other accent varieties, but these results are not reported in this paper.

Procedures

All speakers were given time to familiarize themselves with the script and were provided with guidance during the recording sessions by members of the recording team responsible for directing the recording of TOEFL listening assessment inputs. In a few instances, the initial recordings were too fast, and the speakers made a second recording which were similar to the pace of the United States speaker.

Test takers ($N = 21,726$) were randomly assigned to one of nine conditions. All conditions included two identical TOEFL listening section blocks of three inputs and 17 items as described in the *TOEFL listening test* section. Scores on the 34 items were aggregated for each test taker resulting in a scale of 0 to 34. Reliability, based on Cronbach's coefficient alpha was estimated at .90 for the 34 items. Scores on these 34 items were used as a covariate to ensure that the nine conditions in the study included test takers with comparable listening abilities. One section, a lecture, for the third block was different for each of the nine conditions. Scores for the conversation and one of the lectures in this block are not reported in this study. For the other lecture in this block, test takers in each of the nine conditions heard a different speaker give the lecture. This lecture was about 'drop catch' behavior of seagulls and is described in the *listening comprehension measure* section. Scores on the six items designed to assess comprehension of

this lecture were used to indicate differences in comprehension of the nine speech varieties used in the study. Scores were aggregated for each test taker resulting in a score scale of zero to six. The reliability of these six items, based on Cronbach's coefficient alpha was estimated to be .65. The reliability of these six items was similar to the reliability of other sets of six items which were drawn from one listening input in the first two blocks (range was .56 to .63).

After completing the test, participants were asked to respond to the accent familiarity questionnaire, which was sent out the day after the test was administered. Valid responses of 4,693 (22%) of the test takers were received.¹ Scores on the four items for each test taker's familiarity with British and Australian speakers were then aggregated to provide an overall measure of familiarity with a range of four, *not at all familiar* with an accent to 16, *very familiar* with an accent. Scores of eight and below were categorized as *not familiar* and scores of nine or above were categorized as *familiar*. This cut-point was based on a score distribution in which a large number of students indicated no ('One' on the scale) or little ('Two' on the scale) familiarity with the accents.

Results

Descriptive statistics

The sample size and standard deviations for scores for each of the accent strengths are presented in Table 2, and the means are shown in Figure 1. As shown in Figure 1, the highest average score was for test takers who listened to the United States accent. Moreover, as strength of accent on the listening stimulus increased, in general, comprehension decreased. The one notable exception to this trend was for the speaker with a 2.6 accent strength; the mean score for test takers who listened to this accent was higher than the trend would predict.

¹ Unfortunately, it was not possible to ask test takers to complete the questionnaire during or immediately after completing the assessment. The best possible solution was to send out a request to test takers to complete the survey online. This resulted in a less than desirable response rate.

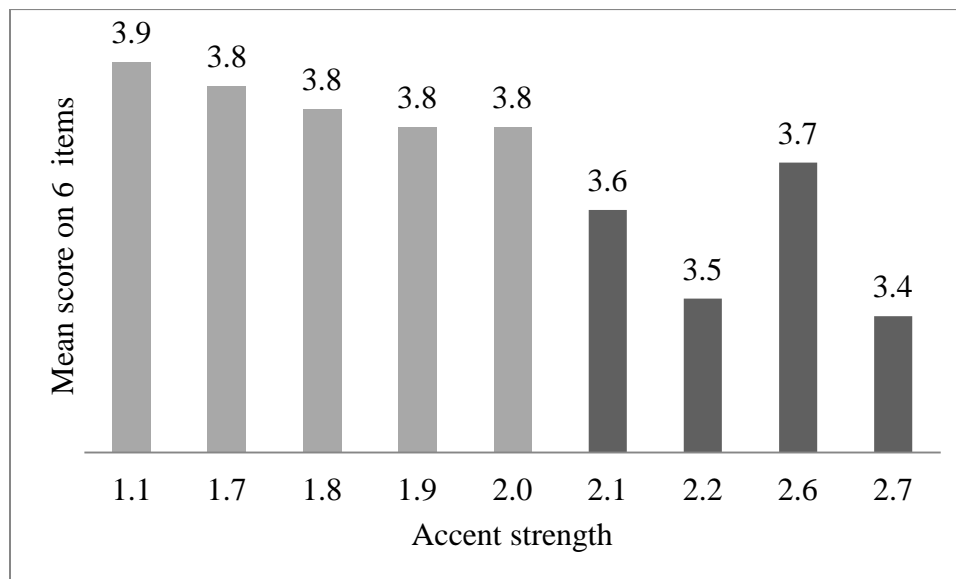
Table 2

Descriptive statistics of scores for each of the nine speech variety conditions

Accent Strength	N	Mean	SD
1.1	4,362	3.9	1.7
1.7	2,174	3.8	1.7
1.8	2,135	3.8	1.7
1.9	2,160	3.8	1.7
2.0	2,168	3.8	1.7
2.1	2,154	3.6	1.8
2.2	2,168	3.5	1.7
2.6	2,214	3.7	1.7
2.7	2,191	3.4	1.7
Total	21,726	3.7	1.7

Figure 1

Scores for each of the nine speech variety conditions



Relationship between comprehension and strength of accent

To answer the first research question which aimed to determine the relationship between an L2 listener's comprehension and the strength of a speaker's accent, a one-way analysis of variance (ANOVA) was conducted. To make the results more comparable to some of the previous studies, it was deemed important to conduct a general analysis that did not control for familiarity and proficiency within a condition. Moreover, each condition had over 2,000 test takers based on random assignment, making it likely that listening proficiency within groups was comparable. Thus, a covariate for listening ability was not used in this analysis. A fixed effects model was used, in which strength of accent with nine conditions was the independent variable and aggregated scores on the six multiple choice items was the dependent variable. An omnibus F test indicated a significant effect for strength of accent, $F_{(8, 21,716)} = 20.20, p < .05, \eta^2 = .01$. A post hoc comparison of means using Dunnett's test at $\alpha = .05$ indicated that listeners who heard speakers with accents of 2.1 and stronger attained scores which were significantly lower than those who listened to the United States speaker. Listeners who heard accents with strengths of 2.0 or weaker did not attain scores that were significantly different than those who listened to the United States speaker. Effect sizes for significantly different scores were: 2.1, Cohen's $d = .15$; 2.2, Cohen's $d = .24$; 2.6, Cohen's $d = .10$, and 2.7, Cohen's $d = .25$.

Relationship among familiarity with and strength of accent and listening comprehension

To determine the relationship among an L2 listener's comprehension and familiarity with a speaker's accent and the speaker's strength of accent, Australian and British accents were considered separately. It was felt that two separate analyses would make it more possible to compare the results of this study to others, and familiarity with the two varieties might be conceived of somewhat differently. Two separate analogous analyses were conducted, both of

which included a covariate for listening ability, because N sizes were rather small for some conditions.

Descriptive statistics for the analysis which examined the effects of familiarity with and strength of Australian accents on listening comprehension are provided in Table 3. As shown, a relatively small number of test takers reported familiarity with Australian accents.

Table 3

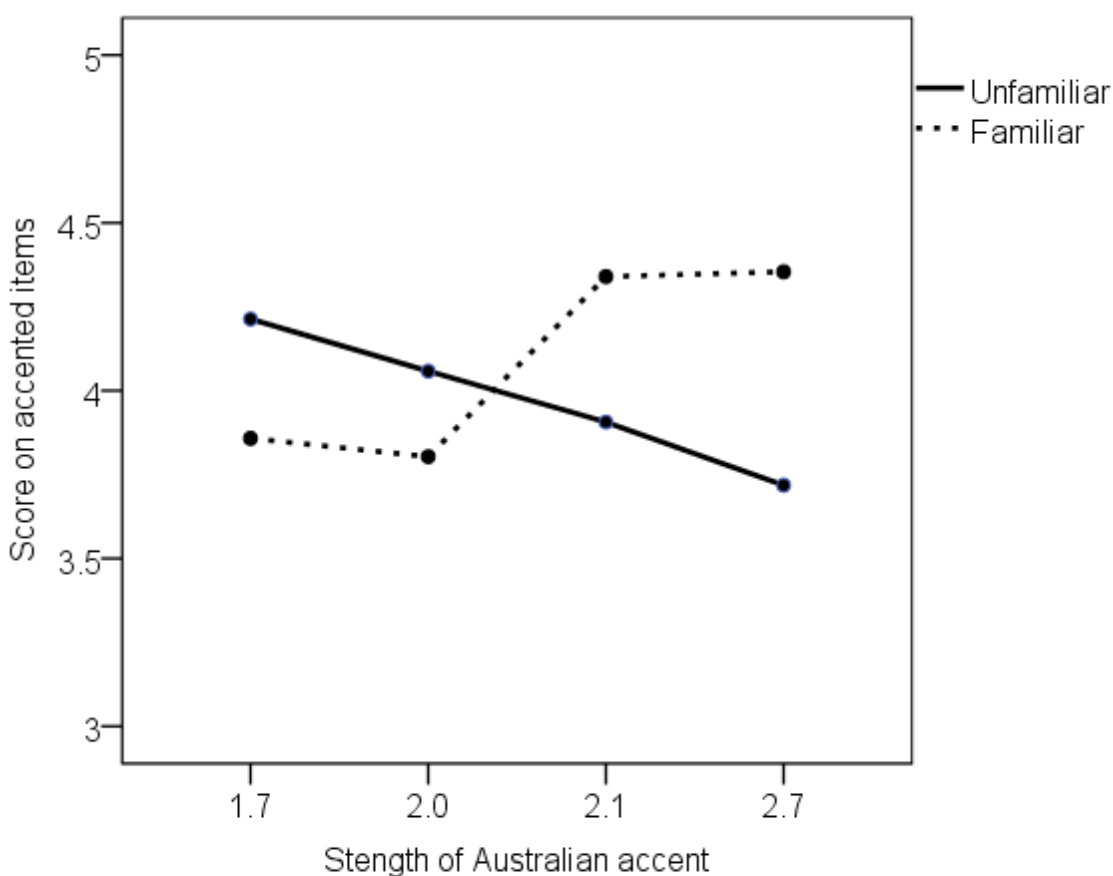
Descriptive statistics for strength of and familiarity with Australian accent

Accent Strength	Familiarity	N	Mean	SD
1.7	Not familiar	461	4.2	1.6
	Familiar	43	4.1	1.9
2.0	Not familiar	408	4.1	1.7
	Familiar	60	3.9	1.6
2.1	Not familiar	417	3.9	1.6
	Familiar	49	4.4	1.7
2.7	Not familiar	425	3.7	1.7
	Familiar	39	4.2	1.7
Total		1,902	4.0	1.7

A two-way ANCOVA was conducted with strength of accent (four treatment levels) and familiarity (two treatment levels) with Australian accents as independent variables, listening ability as measured by the 34-item TOEFL listening assessment as a covariate, and score on the six items based on the lecture delivered by the different speakers as the dependent variable. A significant interaction between strength of accent and familiarity with Australian accents was found: $F_{(3, 1894)} = 6.05, p < .05$, partial $\eta^2 = .01$. Given a disordinal interaction, main effects are not reported (Keppel and Wickens 2004). The interaction is presented in Figure 2.

Figure 2.

Relationship between strength of and familiarity with Australian accents and a listener's comprehension.



As shown in Figure 2, for test takers who are familiar with Australian accents, depicted with the dotted line, in general, scores on the items increase as familiarity with a speaker's accent increases. On the other hand, for test takers who are not familiar with Australian accents, depicted with the solid line, scores decrease as the strength of a speaker's accent increases.

Descriptive statistics for the analysis which examined the effects of familiarity with and strength of British accents on listening comprehension are provided in Table 4.

Table 4

Descriptive statistics for strength of and familiarity with British accents

Accent Strength	Familiarity	N	Mean	SD
1.8	Not familiar	170	3.9	1.6
	Familiar	284	4.3	1.6
1.9	Not familiar	175	3.9	1.6
	Familiar	300	4.2	1.6
2.2	Not familiar	179	3.2	1.7
	Familiar	265	4.1	1.6
2.6	Not familiar	195	3.6	1.8
	Familiar	289	4.3	1.6
Total		1,857	4.0	1.6

To understand the relationship among strength of and familiarity with British accent on listening comprehension, a two-way ANCOVA was conducted with strength of accent (four treatment levels) and familiarity (two treatment levels) with British accents as independent variables, listening ability as measured by the 34-item TOEFL listening assessment as a covariate, and score on the six items based on the lecture delivered by the different speakers as the dependent variable. No significant interaction between strength of and familiarity with accent was found, $F_{(3, 1846)} = 1.69$, ns. However, a main effect for strength of accent $F(3, 1846) = 10.99$, $p < .05$, partial $\eta^2 = .02$ and a main effect for familiarity with accent, $F(1, 1846) = 16.09$, $p < .05$, partial $\eta^2 = .01$ were observed. Given that a main effect for accent strength was investigated in the *Relationship between comprehension and strength of accent* section, post hoc tests were not conducted to compare the effects of strength of accent and were not needed for the dichotomous comparison of effects of familiarity. Test takers who were familiar with British accents achieved significantly higher scores than those who were not familiar with British accents independent of

strength of accent. Moreover, as was found in the first analysis, as strength of accent increased, listening scores tended to decrease independent of familiarity with accent.

Discussion

Relationship between strength of accent and listening comprehension

The first research question aimed to determine the degree to which strength of accent is related to listening scores. The results indicated that as strength of accent increased, listening scores decreased. This effect became significant on scores at the point when the accent moved beyond noticeable, defined as an accent that was stronger than ‘2’ on the Strength of Accent Scale. These effect sizes were quite small, but given the mildness of the accents used in the study, it is reasonable to conclude that these effects should not be discounted. Moreover, given the general pattern of decline between strength of accent and listening scores, it is likely that accents stronger than those used in the study would be associated with even lower listening scores. It is quite interesting to find that accents judged to be only slightly more than *noticeably different than I am used to but did not require me to concentrate on listening more than usual* and much less than *was noticeably different than what I am used to and did require me to concentrate on listening more than normal without decreasing understanding* did affect comprehension. For example, only 7 of the 100 judges felt that they had decreased understanding of the British male speaker who was rated 2.2 overall, and well over half of the listeners did not feel the accent was more than *noticeably different than I am used to, but did not require me to concentrate on listening more than usual*. However, listeners who heard this speaker got significantly lower scores than those assigned to listen to the US speaker. These findings suggest that listeners may overestimate the capacity to understand an unfamiliar accent and more importantly that the

comprehension of some listeners can be impacted by accents judged to be quite similar to the familiar variety.

Although a strong negative trend between strength of accent and scores on the listening items were found, one notable exception existed in the data. The British female, rated by the 100 academic judges as 2.6 on the Strength of Accent Scale, was not as difficult to comprehend as two speakers judged to have accents more similar to the local variety. Exploration aimed at identifying an explanation for this finding suggested that the judgments of the two 20-second audio clips used to judge the strength of this speaker's accent was quite different when compared to the ratings based on the two audio clips of the other speakers in the study (see Table 1). It should be noted that had only the rating of 2.2 been used in the study, the relationship between strength of accent and listening comprehension would have been completely consistent; as strength of accent increases, listening comprehension decreases. Thus, it may well be that the rating of 2.9 was inaccurate. In spite of careful controls in the study, the speech sample that was rated 2.9 may have been judged harsher than it should, for example, due to pronunciation of a particular lexical item. This underscores the importance of carefully ensuring that speech samples are appropriate when rating accent strength. All factors, which could conceivably affect a rater's judgment of a speech sample, such as pace, pausing, vocabulary, and speech segment context should be controlled when accent strength is judged. This may help to limit misjudgments based on the Strength of Accent rating scale, which is a likely explanation for the rating of 2.9 of the British female speaker who had an average rating of 2.6.

Relationship between familiarity with and strength of accent and listening comprehension

Given that strength of accent was found to affect listening comprehension, the second research question, which related to the extent to which familiarity, strength of accent, and listening

comprehension are related, had increased importance. As noted in the results section, the effects of Australian and British accents were analyzed separately. The results indicated a slightly different relationship among these variables for Australian and British speakers. Test takers who were familiar with Australian accents received increasingly higher scores with increasingly stronger Australian accents, while test takers who were not familiar with Australian accents attained increasingly lower scores as accents increased in strength. The size of this effect was rather small, only about one percent of the score variance. In addition, the sample size for the test takers familiar with Australian accents was quite small, decreasing the confidence that could be given to the result. These limitations may help to explain the rather unexpected outcome that test takers familiar with Australian accents did slightly worse than test takers who were unfamiliar with Australian accents for the speakers with the lightest Australian accents. Nonetheless, the general pattern of an increasing disadvantage for test takers unfamiliar with Australian accents as strength of accent increased suggests that these results should not be ignored. The findings of this study are in line with those of Major et al. (2002), Adank et al. (2009), and Harding (2011), who found a similar accent familiarity advantage. This finding coupled with the findings of previous research provides rather strong support for the commonsensical contention that familiarity with accent does provide a listening comprehension advantage. These results support Buck's (2001) position that if it is necessary to use an accent that is unfamiliar to test takers, one should be selected that is equally unfamiliar to all (162).

The relationship among familiarity, strength of accent, and comprehension of British accents was slightly different than for Australian accents. Familiarity with British accents was found to be an advantage in comprehending British speakers, and stronger accents were more difficult to comprehend. The failure to find an interaction between familiarity and strength of

accent may have been due to the problematic judgment of the British female accent that was judged to have an accent with a strength of 2.6. Of course, it is also possible that these two contexts are quite different and the relationship among these variables is different.

It is possible that test takers' perceptions of their familiarity with Australian and British speech varieties may have impacted the findings of the study. Test takers may believe that they are more familiar with British speakers because they are more commonly heard in the United States and many other parts of the world through media sources such as the BBC. On the other hand, test takers may believe that they are generally not familiar with Australian accents because they are not as commonly heard in the United States. Well over half of the test takers indicated familiarity with British accents, whereas less than 10% reported familiarity with Australian accents¹. While there are other possible explanations, it is plausible that the few test takers who reported familiarity with Australian accents had a high degree of familiarity with these accents while those reporting familiarity with British accents may have had less familiarity. Thus, it might be expected that the effect of familiarity would have been stronger for the Australian speakers than the British speakers. If this were the case, it would suggest that familiarity with an accent has a stronger effect on listening comprehension than was found for British accents in the study.

Generalizability and limitations of findings

It is important that the findings of this study are not over generalized. First, only one type of lecture was used. This lecture had certain characteristics which might limit the generalizability of the findings. For instance, these lectures were monologic; conversation or dialogues between two people were not used and may not support these findings. Similarly, the lecture was decidedly

scientific. Other types of lectures, such as ones with humanities content, may not have led to the same results. Second, results may be limited by the degree of accent of the speakers in the lecture. Only rather mild accents were used because it was assumed if mild accents affected comprehension, strong accents would also. This assumption may not necessarily follow, however. Third, accent is only one aspect of a speaker's dialect that could impact test scores. This study controlled for factors such as differences in vocabulary and grammar among Australian, British, and United States speakers by having all speakers read a script written by users of the United States speech variety. Just as accents differ systematically among speakers of particular varieties of English, grammar and vocabulary vary systematically. Thus, it should not be assumed that factors such as these, which were controlled in this study, would not impact test scores if inputs were based on a speaker's own language variety. Fourth, reported familiarity with an accent may not have meant that listeners were familiar with the actual accents of the speakers in the study. The speakers' accents were identified as Australian or British, and listeners were asked to indicate their degree of familiarity with these speech varieties. However, variations of these speech varieties do exist, and therefore, listeners' indications that they were generally familiar with British or Australian accents might not suggest that they were familiar with the specific accents that they heard. Fifth, it should not be assumed that a test with a large number of inputs delivered by speakers with various accents would not affect the results of the assessment, even if all speakers had accents with measures below '2' on the accent measure. The effect of accent on listening comprehension in this study was based on six items. It is possible that accents which are lighter than '2' on the accent scale might have an impact on scores if a large number of items were used. That is, a cumulative effect may occur by having multiple inputs accompanied by a large number of items, even if the accents are quite mild. On the other

hand, research indicates that comprehension of an unfamiliar accent increases in relation to exposure (Clark and Garrett, 2004; Adank and Janse, 2010). As a result, it may be that with more than one stimulus from the same unfamiliar accent, comprehension might increase as the test taker progresses through the test. Further research which disentangles these effects for a test of a particular length is needed. Finally, while the test taker population was large and diverse, it only included participants who took TOEFL iBT. It could therefore not be assumed that other test taker populations would be impacted by the strength of the accents in this study in the same way.

Implications

The implications are founded on the premise that agreement could be reached on an English speech variety to which others could be compared. While such a variety was justified in this study, given that the test had previously only included one speech variety, this may not hold for other contexts. In such cases, stakeholders will need to identify an acceptable variety, and in an international context, agreement on such a variety has not been reached (Seidlhofer 2003).

The current study suggests that a defensible measure of accent can be used to judge the strength of a speaker's accent. Ratings on the Strength of Accent Scale provided a good indication of the degree to which listening comprehension would be impacted by the accent of a speaker. It is important to note, however, that this measure may not have provided an accurate estimate of the strength of accent for one of the speakers in the study, suggesting the need for a wider range of speakers and further research before strong claims can be made about the effectiveness of the instrument. An important question that arises from this study relates to the number of judges that would be needed to reliably rate a speaker's strength of accent. While 100

were used in this study, it is conceivable that a much smaller number, such as five or six would be sufficient to achieve a reasonable level of reliability.

Only speakers from a small number of English varieties were included in this study, but it can be argued that these results suggest that the Strength of Accent Scale could be used to judge the strength of any accent. As has been discussed, accents with strengths which did not exceed '2' would be unlikely to significantly impact listening comprehension scores. This may make it possible to sample from any variety of accents, so long as its strength does not exceed '2' on the Strength of Accent Scale. Thus, this research suggests a possible direction for those who desire to assess a more multidialectal construct of listening comprehension without unfairly impacting the scores of some of the test takers. To adequately assess L2 listening comprehension, it is important that the varieties of English that are frequently encountered in the target language use domain be included on test inputs. Just as tests which do not include listening inputs spoken by speakers of the most common English variety from the domain of generalization are likely to lead to scores which are not representative of a test taker's listening ability, limiting inputs to speakers with only one dialect from the domain of generalizations likely leads to scores which are not completely valid indicators of a test taker's listening ability (Harding 2011; Taylor and Geranpayeh 2011; Abeywickrama 2013). On the other hand, to ensure test fairness, it is essential that test takers do not encounter accents on the assessments which unfairly give one test taker an advantage over another. If one test taker does better than another simply because the accent presented on the assessment is one with which the test taker happens to be familiar, this would be a threat to test fairness. Use of a measure of strength of accent to select speakers for an assessment makes it possible to include a variety of mild accents without unfairly disadvantaging certain test takers. It is also likely that such an approach would lead to positive washback

because test takers would try to develop a multidialectal listening ability. This may suggest that as test takers develop familiarity with various accents, over time, more diverse accents could be used without unfairly disadvantaging some test takers.

Conclusion

This study sheds light on what it means to broaden the construct of listening comprehension to become more multidialectal. The results suggest that accents with strengths which are perceived to require extra effort from some listeners for full comprehension inhibit comprehension. The findings also suggest a listening comprehension advantage for test takers who are familiar with accents. Given these findings, it would be unfair to test takers and professionally irresponsible to use unmeasured accents for listening comprehension assessment inputs. However, it is also a threat to the validity of an assessment to use only one variety of English to assess English listening comprehension when it is apparent that more than one variety of speech is commonly encountered in the target language use domain. Given this dilemma, the researchers propose that to make the construct of listening comprehension more multidialectal, various accents which can be shown to not unfairly impact scores, be included. This can be accomplished by devising a valid measure of accent and using it to select speakers. Such an approach may lead to a gradual broadening of the listening construct, one that better represents the listening situations that exist in the real world while maintaining an assessment that does not unfairly disadvantage some test takers. The construct of listening comprehension must be expanded to include more than one preferred accent, but as Elder and Harding (2008) insist, ‘Attempts by language testers to address the challenge of EIL (English as an International Language) must proceed in an evidence-based and consultative manner’ (34.2).

¹ As a comparison, more than 90% of TOEFL iBT test takers indicate that they are familiar with US accents.

References

- Abeywickrama, P.** 2013. 'Why not non-native varieties of English as listening comprehension test input?' *RELC Journal* 44/1: 59-74.
- Adank, P., B. Evans, J. Stuart-Smith, and S. Scott.** 2009. 'Comprehension of familiar and unfamiliar native accents under adverse listening conditions.' *Journal of Experimental Psychology* 35/2: 520-529.
- Adank, P. and E. Janse.** 2010. 'Comprehension of a novel accent by young and older listeners.' *Psychology and Aging* 25/3: 736-740.
- Anderson-Hsieh, J. and K. Kohler.** 1988. 'The effect of foreign accent and speaking rate on native speaker comprehension.' *Language Learning*, 38/4: 561-613.
- Bachman, L. F., & Palmer, A. S. (1996). Language Testing in practice. Oxford University Press.**
- Bilbow, G. T.** 1989. 'Towards an understanding of overseas students' difficulties in lectures: A phenomenographic approach.' *Journal of Further and Higher Education* 13: 85-89.
- Bradlow, A. R. and T. Bent.** 2008. 'Perceptual adaptation to non-native speech.' *Cognition* 106/2: 707-729.
- Buck, G.** 2001. *Assessing Listening*. Cambridge University Press.
- Clarke, C. M. and M. F. Garrett.** 2004. 'Rapid adaptation to foreign accented English.' *Journal of the Acoustical Society of America* 116: 3647-3658.
- Department of Institutional Research.** 2007. Faculty by ethnic group [Online]. <http://www.irs.ttu.edu/NEWFACTBOOK/Faculty/2007/F07ETHNIC.htm>
- Derwing, T. M. and M. J. Munro.** 1997. 'Accent, intelligibility, and comprehensibility: Evidence from four L1s.' *Studies in Second Language Acquisition*. 20: 1-16.

- Derwing, T. M. and M. J. Munro.** 2009. 'Putting accent in its place: Rethinking obstacles to communication.' *Language Teaching* 42/4: 476-490.
- Eisentein, M. R. and D. Berkowitz.** 1981. 'The effect of phonological variation on adult learner comprehension.' *Studies in Second Language Acquisition* 4: 75–80.
- Ekong, P.** 1982. 'On the use of an indigenous model for teaching English in Nigeria.' *World Language English* 1: 87–92.
- Elder, C. and A. Davies.** 2006. 'Assessing English as a lingua franca.' *Annual Review of Applied Linguistics* 26: 282-301.
- Elder, C. and L. Harding.** 2008. 'Language testing and English and an international language: Constraints and contributions' in Sharifian, F. and M. Clyne (eds.): *Australian Review of Applied Linguistics* (special forum issue) 31/3: 34.1–34.11.
- Field, J.** 2004. 'Pronunciation acquisition and the individual learner.' Presentation at the IATEFL Joint Pronunciation and Learner Independence Special Interest Groups Event, University of Reading, 26 June 2004.
- Gass, S. and E. M. Varonis.** 1984. 'The effect of familiarity on the comprehensibility of nonnative speech.' *Language Learning* 34/1: 65-89.
- Harding, L.** 2011. *Accent and Listening Assessment: A Validation of the Use of Speakers with L2 Accents on an Academic English Listening Test*. Peter Lang.
- Keppel, G. and T. Wickens.** 2004. *Design and Analysis: A Researcher's Handbook, Fourth Edition*. Pearson-Prentice Hall.
- Major, R., S. Fitzmaurice, F. Bunta, and C. Balasubramanian.** 2002. 'The effects of nonnative accents on listening comprehension: Implications for ESL assessment.' *TESOL Quarterly* 36/2: 173-190.

- Munro, M. J., and T. M. Derwing.** 1995a. 'Foreign accent, comprehensibility and intelligibility in the speech of second language learners.' *Language Learning* 45: 73-97.
- Munro, M. J. and T. M. Derwing.** 1995b. 'Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech.' *Language and Speech* 38/3: 289-306.
- Ortmeyer, C. and J. Boyle.** 1985. 'The effect of accent differences on comprehension.' *RELC Journal* 16/2: 48-53.
- Pennington, M. C.** 1996. *Phonology in English Language Teaching*. Longman.
- Schmid, P. and G. Yeni-Komshian.** 1999. 'The effects of speaker accent and target predictability on perception of mispronunciation.' *Journal of Speech, Language, and Hearing Research* 42: 56-64.
- Seidlhofer, B.** 2003. A Concept of International English and Related Issues: From 'Real English' to 'Realistic English?' in Language Policy Division, Council of Europe, Strasbourg.
- Smith, L. and J. Bisazza** 1982 . 'The comprehensibility of three varieties of English for college students in seven countries.' *Language Learning* 32/2: 259-269.
- Tauroza, S. and J. Luk.** 1997. 'Accent and second language listening comprehension.' *RELC Journal* 28: 54-71.
- Taylor, L.** 2006. 'The changing landscape of English: Implications for English language assessment.' *ELT Journal* 60/1: 51-60.
- Taylor, L. and A. Garenpayeh.** 2011. 'Assessing English for academic purposes: Defining and operationalizing the test construct.' *Journal of English for Academic Purposes* 10: 89-101.

Appendix A

Strength of Accent Scale

1. The speaker's accent was **NOT** noticeably different than what I am used to and did **NOT** require me to concentrate on listening any more than usual. The accent did **NOT** decrease my understanding.
2. The speaker's accent was noticeably different than what I am used to but did **NOT** require me to concentrate on listening any more than usual. The accent did **NOT** decrease my understanding.
3. The speaker's accent was noticeably different than what I am used to and did require me to concentrate on listening more than usual. However, the accent did **NOT** decrease my understanding.
4. The speaker's accent was noticeably different than what I am used to and did require me to concentrate on listening more than usual. The accent slightly decreased my understanding.
5. The speaker's accent was noticeably different than what I am used to and did require me to concentrate on listening more than usual. The accent substantially decreased my understanding.

Appendix B

Accent familiarity questionnaire

1. Overall, how familiar are you with the following English accents?

	Not at all familiar	A little familiar	Familiar	Very familiar
• American (US)	1	2	3	4
• Australian	1	2	3	4
• British (UK)	1	2	3	4
• Other native (for example, Canadian, New Zealander)	1	2	3	4
• Non-native English speakers (for example, Chinese, Indian, Spanish or German speakers with clear pronunciation)	1	2	3	4

2. How often do you hear the following English accents on TV, radio or the internet?

	Rarely	Sometimes	Often	Very often
• American (US)	1	2	3	4
• Australian	1	2	3	4
• British (UK)	1	2	3	4
• Other native	1	2	3	4
• Non-native	1	2	3	4

3. How often do you hear the following English accents in face-to-face communication? (Consider communication with classmates, friends, colleagues, teachers, and others who have these accents).

	Rarely	Sometimes	Often	Very often
• American (US)	1	2	3	4
• Australian	1	2	3	4
• British (UK)	1	2	3	4
• Other native	1	2	3	4
• Non-native	1	2	3	4

4. How long have you studied English with teachers who have the following accents?

	Not at all	less than 1 year	1-2 years	more than 2 years
• American (US)	Not at all	less than 1 year	1-2 years	more than 2 years
• Australian	Not at all	less than 1 year	1-2 years	more than 2 years
• British (UK)	Not at all	less than 1 year	1-2 years	more than 2 years
• Other native	Not at all	less than 1 year	1-2 years	more than 2 years
• Non-native	Not at all	less than 1 year	1-2 years	more than 2 years