The effect of prompt accent on elicited imitation assessments in English as a second language

Jacob G. Barrows & Troy L. Cox
Brigham Young University, USA

Elicited imitation (EI) assessments have been shown to discriminate well between speakers across proficiency levels, but little has been reported on the effect L2 accent has on test-takers' ability to understand and process the test items they hear. Furthermore, no study has investigated the effect of accent on test-taker perceptions of EI tests. This study examined the relationships among accent, accent familiarity, EI test item difficulty and test scores. To investigate, self-reports of students' exposure to different varieties of English were obtained from a pre-assessment survey. An EI test (63 items) was then administered in which English language learners (n = 213) in the United States listened to test items in three varieties of English: American English, Australian English, and British English. A Rasch analysis found that the test had high reliability (person separation = .94), with intended item level and accent both having a significant effect on test item difficulty. Survey results indicated a moderate relationship between an examinee's familiarity with a particular accent and their person ability estimate measures. These findings suggest that prompt accent should be considered in EI test development.

Keywords: Elicited imitation, accent, listening comprehension, speaking assessment, Rasch measurement

Introduction

Globalization, modern communication, and media technology are bringing new challenges to the field of language assessment. This phenomenon is especially true of English, which has developed robust L1 (native) and L2 (non-native) varieties, both regionally and nationally. Although it enjoys the privileged status of being studied and spoken internationally in many different contexts, learners might only be exposed to one or two varieties. An EFL learner in Mexico, for example, might have more opportunities

Email address for correspondence: troyc@byu.edu

to be exposed to American English than to British English through either media or personal interactions. How, then, would an ESL assessment with a listening component accurately measure that student's ability if the prompts contained a British variety of English? Those who design ESL tests should be cautious of potential difficulties that can arise when making broad assumptions concerning a learner's background with any given variety of English.

While an ESL learner's familiarity with different language varieties can be a challenge for all types of assessments, speaking and listening assessments are particularly problematic. These tests typically utilize audio prompts, and any audio prompt is, by necessity, colored by the speaker's accent. To address this challenge, this research seeks to better understand the interaction between a learner's familiarity with regional varieties of English and their results on an elicited imitation (EI) assessment. The design of EI assessments is fairly simple: students listen to an audio prompt and attempt to repeat, verbatim, what they hear (Yan et al., 2016). Their repetitions are recorded and graded for accuracy. Yan et al. (2016) found in a meta-analysis of 21 EI studies (n = 1,089) that "EI tasks in general have a strong ability to discriminate between speakers across proficiency levels" (p. 498) and have been found to be highly reliable.

One challenge of EI testing is test-taker perception, as some test-takers have difficulty seeing how a sentence-repetition task accurately measures language proficiency (Graham et al., 2008; Moulton, 2012; Van Moere, 2012; Vinther, 2002). Anecdotes from students suggest that this problem is compounded when test-takers listen to EI prompts in an accent they are less familiar with. These negative test-taker perceptions may hinder test-taker motivation and, therefore, test performance (Chan et al., 1997).

While research has been conducted on the effect of accent on listening comprehension (Kang et al., 2019; Harding, 2018; Ockey et al., 2016), little has been done to examine the extent to which this holds true with EI. With EI, instead of simply indicating that the content of a prompt was correctly understood, test-takers respond to stimuli in a way that indicates correct parsing and processing of each morphological and syntactical aspect of the prompt. Thus, prompt accent potentially has more impact on EI performance than other listening comprehension measures.

Background

Elicited imitation (EI) assessment

In EI assessments, test-takers are provided with a prompt that they must attempt to repeat verbatim. In many assessments, including the one used in this study, the test

administration is automated, with test-takers listening to audio recordings and repeating the utterances into a microphone. The recorded utterances are later scored by marking errors in the repetition (e.g., the syllables incorrectly repeated, words omitted, etc.).

The nature of EI test administration and scoring offers a number of advantages (technology permitting), among which time and cost effectiveness are most notable. When it is automated, an EI test can be administered to several students at the same time under the supervision of an administrator, even one who is not highly trained in the task (Graham et al., 2008). Test scoring can also be quick and objective (Matsushita & Lonsdale, 2012) and, when scored by humans, does not require extensive training to rate speech accurately (Millard & Lonsdale, 2014). Future developments in automated speech recognition (ASR) technology may even further reduce the cost and increase the reliability of EI assessment (Cox et al., 2015).

Though speech is the object that is scored in EI, it is generally accepted that EI assessment reflects broad language proficiency or implicit grammatical knowledge (Yan et al., 2016). However, a test-taker's listening ability is the gatekeeper to measuring his or her language proficiency with EI. Cox and Davies (2012) compared an EI assessment with a number of other proficiency measures (i.e. a speaking proficiency interview, a writing placement exam, and a computer adapted exam of listening, reading, and grammar) and found that although there were correlations between EI and the other assessments, EI had the highest correlation with the listening exam (r = .74).

Listening and accents

Of the studies investigating the interaction between listening comprehension and different accents, nearly all have found that an unfamiliar accent hinders listening. The same results have been found with different types of accents, including L2-accented English (Anderson-Hsieh & Koehler, 1988; Brunfaut & Révész, 2015; Clarke & Garrett, 2004; Gass & Varonis, 1984; Harding, 2011; Varonis & Gass, 1982), regional and international accents for L1 listeners (Adank & McQueen, 2007; Adank et al., 2009; Floccia et al, 2006; Major et al., 2005), regional and international accents for L2 listeners (Kang et al., 2019; Major et al., 2005; Ockey & French, 2014; Ockey, et al., 2016), L1 ethnic accents (Major et al., 2005), and even artificial accents (Maye et al., 2008; Wingstedt & Schulman, 1984).

Research attempting to examine accent must apply a clear and defensible definition of the word *accent*. Clearly defining accent is a difficult task, however, as can be seen by the number of studies that fail to do so (e.g., Abeywickrama, 2013; Clarke & Garrett, 2004; Gass & Varonis, 1984). One vein of thought is that an accent is an attribute of speech that varies according to geography, native language (when nonnative speech is examined),

ethnicity, or individual speaker (Abeywickrama, 2013; Anderson-Hsieh & Koehler, 1988; Clarke & Garrett, 2004). Those who take this perspective often view accent as a unidimensional construct for which a speaker's provenance is the only variable. In these studies, a speaker's identity or linguistic background is usually sufficient evidence of accent, though, in some cases, expert judgment (Adank et al., 2009) or phonetic analysis is conducted to empirically delineate geographical accent boundaries (Adank & McQueen, 2007).

For the current study, this perspective on accent presents a number of challenges, including questions about how a national accent is defined, whether a national standard is representative of what the listeners are likely to have encountered, or whether such standards are actually spoken in Australia, the U.S., and the U.K.—the three countries whose accents are considered here. Particular attention must be paid to the U.K., whose traditional standard, Received Pronunciation, has been on the decline for several years in favor of more regional varieties (Mugglestone, 2007). More importantly, though, the perspective that accent is defined by geography is only focused on the speakers and ignores the listeners' experience.

For this study, a more appropriate definition of accent should account for the complex nature of language as a shared experience between two interlocutors. Likewise, Derwing and Munro (2009) take a different approach and present a method of understanding accent according to listeners' perceptions rather than as an attribute that exists independently as part of speech itself. They define the construct of accent as having three related, yet partially independent, dimensions: *accentedness* is how different one variety sounds from the listener's local variety, *comprehensibility* is how difficult a listener believes a variety is to understand, and *intelligibility* is how much a listener is able to understand.

In the present study, test-takers' perceptions of how difficult the accents are to understand correspond to Derwing and Munro's notion of comprehensibility, and strength of accent corresponds to the degree of accentedness as determined by a panel of judges. To be clear, the notion of *accent* is a broad one and may encompass all of these dimensions, while *accentedness*, as defined by Derwing and Munro (2009) and operationalized in this study, is more precise. It should be noted that the phrase *strength of accent*—which is applied to surveys conducted here and in other studies—is in fact used to measure *accentedness*, not the more general concept of *accent*. Though this term may be confusing, it is employed here for the sake of continuity with other studies that have investigated accentedness in similar ways.

Defining accentedness as a scalar dimension requires researchers to obtain a measurement of it. Derwing and Munro, who pioneered this definition of accentedness,

created a rubric and obtained this measurement by asking the listeners in their experiments to rate the audio samples on a scale (Derwing & Munro, 1997; Munro & Derwing, 1995a, 1995b). These listeners participated in every aspect of the study—both the accent judgment tasks and the listening comprehension tasks—and were native speakers of English. In contrast, Ockey and French (2014) used a separate panel of judges to rate accentedness. These judges were native and highly proficient nonnative speakers. Similar strength of accent scales and panels of judges have also been used in related studies (Major et al., 2002, 2005).

Research questions

EI tests have often been treated with skepticism on the part of test-takers, and there is sometimes a perception that they are not valid (Graham et al., 2008; Moulton, 2012; Van Moere, 2012; Vinther, 2002); this skepticism may be exacerbated when audio prompts contain unfamiliar accents and test-takers perceive that they face an unfair disadvantage in the test. Previous research on test-taker perceptions indicates that negative impressions of test validity may lead to lower motivation in test-takers, which, in turn, may reduce test performance (Chan et al., 1997). Thus, to improve the quality of EI testing, this research aims to answer the following questions:

1. What effect does speaker accent on the prompt have on EI test item difficulty?
2. To what extent does the examinees' familiarity with the accent have on their EI test scores?

Methods

In order to determine the effect that familiarity with regional accents has on EI test scores, and the effect that examinee perception of the accents has on test difficulty, two instruments were created: (1) a pre-test survey that gathered data on participants' experience with and exposure to different varieties of English; and (2) an EI test with recordings of American, Australian, and British speakers. These instruments were then administered to ESL students studying at a large U.S. university. The EI test was subsequently scored by human raters, after which the results were analyzed.

Pre-test survey

Prior to beginning the EI test, participants completed a brief survey about their experience with and exposure to different varieties of English including (1) American English, (2) Australian English, (3) British English, (4) other native English varieties, and

(5) nonnative English varieties. The survey included five questions that ranged from familiarity to context of exposure and can be found in the Appendix.

EI test

A previous test that was designed to assess oral proficiency based on the ACTFL (American Council on the Teaching of Foreign Languages) proficiency guidelines (Cox et al., 2015) provided the framework for this EI test, including the text for most of the items. The selected items had equivalent item difficulty levels, so the only manipulation used was to create new audio recordings with the target accents for those items. It was also essential to find speakers whose accents were of comparable strength. This precaution was to ensure that the results were based on accent variety, not strength of accent.

Selection of speakers

The speakers for these recordings were selected from nine volunteer undergraduate students at a large research university in the U.S. Two were American, two were Australian, and five were British. Each of the volunteers read the full list of sentences in a sound recording booth. Six recordings from each participant—for a total of 54 items—were then inserted into a strength-of-accent survey.

Each item of the strength-of-accent survey asked participants to "identify how different the English sounds were from [their] local variety," after which they were presented with the audio clip and a 7-point rating scale ranging from "not at all different" to "very different." This strength-of-accent scale and the questions used to elicit responses were adapted from previous research (Major et al., 2002, 2005; Ockey & French, 2014; Ockey, 2018). The main purpose of this survey was to select speakers who had relatively similar strength of accent according to listeners of other countries to thus mitigate some of the issues inherent with geography-based surveys. In addition, it had the added benefit of ensuring that the Australian and British volunteers still retained their original accents despite their time in the U.S.

The strength-of-accent survey was administered to 126 participants—42 from each country (i.e. the U.S., Australia and the U.K.)— in a paid online research panel. All participants were native speakers of English, their age ranging from 18 to 30. They had lived in their country of origin for the last 10 or more years. The American participants were located all across the continental U.S.; the Australian participants were mostly from urban centers on the east coast of Australia; and almost all British participants were located in England—two were in Scotland and one was in Wales.

Survey results identified speakers from each country who had similar strengths of accent when judged by foreign listeners (foreign-accent rating) as well as when judged by

listeners from their own countries (own-accent rating) (see Table 1). Speaker 2 was selected as the American speaker even though Speaker 1 (American) came closer to Speakers 3 (Australian) and 5 (British) in own-accent ratings for two reasons. First, it allowed for the lowest own-accent rating for each accent to be selected, but more importantly Speaker 2 was the same gender as Speakers 3 and 5, and thus controlled for gender.

Table 1. Strength of accent rating of speakers who volunteered for this study

Speaker		Mean accent rating by origin of rater				
ID	Origin	US	Australia	UK	Foreign total	Own accent
1	US	2.0	4.6	5.5	5.1	2.0
***2**	**US**	**1.2**	**4.4**	**5.5**	**5.0**	**1.2**
***3**	**Australia**	**4.0**	**1.8**	**5.1**	**4.5**	**1.8**
4	Australia	4.3	2.6	4.8	4.6	2.6
***5**	**UK**	**5.4**	**4.6**	**2.9**	**5.0**	**2.9**
6	UK	4.4	3.1	3.3	3.8	3.3
7	UK	5.2	5.2	3.3	5.2	3.3
8	UK	4.1	4.2	3.7	4.2	3.7
9	UK	6.0	5.6	3.4	5.8	3.4

1 = "Not at all different", 7 = "Very different"

Note: Asterisked, bolded, shaded rows indicate speakers selected for the study.

Test form design

The EI test was designed using Wilson's (2005) construct map as the model. The items were written to represent three intended ACTFL item levels which roughly correspond to the following CEFR levels: Intermediate (A2), Advanced (B2), and Superior (C2) (see Figure 1). To represent the different levels, the criteria of length, vocabulary frequency, and grammatical complexity were conjointly manipulated so that the Superior items were the longest, had the least frequently used vocabulary and most grammatical complexity while the Intermediate were the shortest with the most frequently used vocabulary and least grammatical complexity.

The items were then distributed to test forms and were composed of two interwoven components: anchoring items (k = 18) and unique items (k = 45) at three different levels. These items were presented in 3 subtests that consisted of 21 items (anchor k = 6, unique k = 15) to examinees in a Latin square design so that the accents would be presented in all orders (see Figure 2). The unique items were always presented in the same order, but the anchor items would move around depending on the test form they were in. This was to control for the possibility that exposure to one accent would prime the listener for the others.

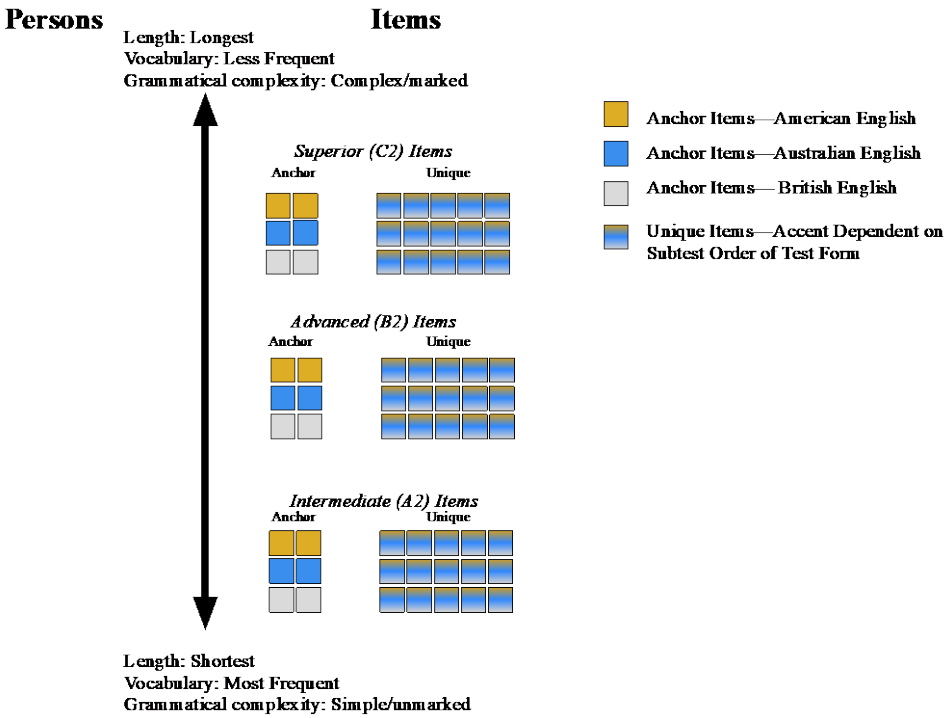

Figure 1. Construct Map of EI Items

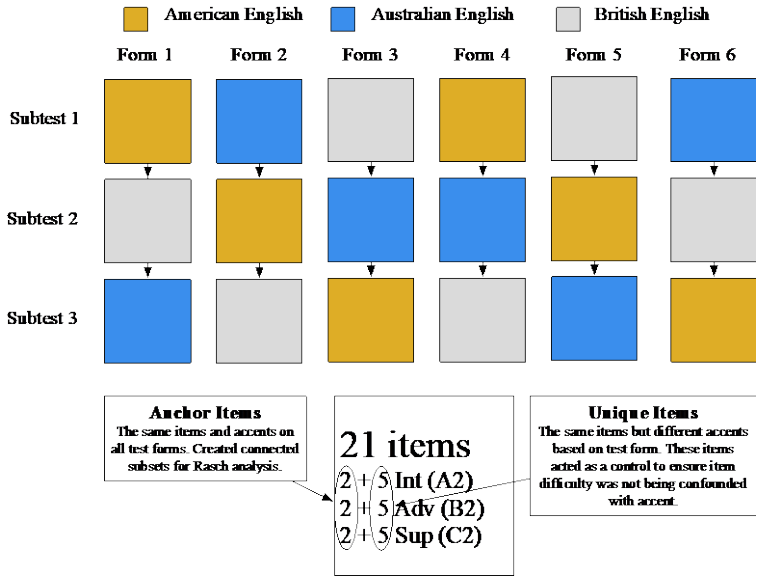

Figure 2. Diagram of intended EI test forms

The anchor items were designed so that all participants were exposed to some items in common (all hearing identical recordings). The unique items, on the other hand, were designed to ensure that each group of participants heard an equivalent amount of each accent but for different items. This was to control for item effect in which the specific features of the EI item might contribute to the score variance more than the accent used.

Originally, these unique items were intended to be divided evenly between intermediate, advanced, and superior levels on each subtest, but an error in the computer programming assigned all intermediate unique items to Subtest 1, all advanced unique items to Subtest 2, and all superior unique items to Subtest 3 (see Figure 3). The items were still inserted into the various test forms for the different groups according to the original design, with the result that all unique items for a given accent were of a single difficulty level. This error was not noticed until after data collection and is considered in the final analysis of the data.

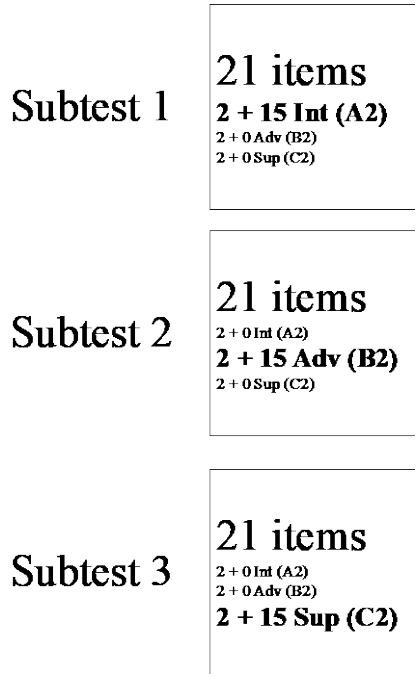

Figure 3. Diagram of actual subtest composition

If the total score had been analyzed using classical test theory, it would have been problematic; however, we used Many-Facet Rasch Measurement (MFRM). MFRM extends the basic Rasch model of examinees and items by incorporating more variables or facets, in this instance, accent. One feature of MFRM is that item difficulty parameters are person independent (McNamara et al., 2019). In other words, the difficulty of specific items is calculated probabilistically, and the relative location of the item difficulty parameter functions independently of the examinees. When each test form has a set of common items—or anchor items—then it does not matter if some of the items unique to each test form vary in degree of difficulty; the item difficulty parameter can still be computed, and the data is still usable.

The anchor and unique items were combined and organized into three parts according to the accent of each prompt, with anchor items and unique items of a given accent occurring in tandem in single blocks of items. The six anchor items preceded the unique items in each block. Organizing the items into blocks was for organizational purposes

only, and test takers did not experience any pauses between blocks or any other indicator (other than a change of speaker) that the test was organized in such a fashion.

To account for ordering effects of prompt accent, six separate forms of the test were created and administered simultaneously. While the content (i.e. the text) of the prompts was identical for each form, the order in which items appeared was different, as was the accent in which unique items were recorded. A one-way ANOVA was performed with the dependent variable (DV) being the Examinee Fair Average generated by the MFRM analysis and the independent variable (IV) being the test form to which the participants were assigned. The analysis indicated no significant difference [$F(5, 207) = 1.72$, $p = .132$]. Therefore, the groups were treated as equivalent with accent order not having an effect.

Table 2. Demographic Information of Research Participants

	Number	Percentage
Gender		
Male	91	42.7%
Female	122	57.3%
Age		
18-25	139	65.3%
26-30	40	18.8%
31+	34	16.0%
Native Language		
Spanish	117	54.9 %
Chinese	26	12.2 %
Portuguese	24	11.3 %
Korean	18	8.5 %
Japanese	14	6.6 %
Russian	7	3.3 %
Other	7	3.3 %

Participants

The study included 213 students at the university's intensive English program that places students in classes that range from beginners to university-ready. The student body was, by design, diverse in language ability and L1 background (see Table 2). The diversity of these students provided a good sample for this study; many students had received significant exposure to and instruction in varieties of English other than American before coming to the U.S.

Instrument administration

The pre-test survey and EI test were administered in the intensive English program's

computer lab under the supervision of trained proctors as part of their regular exit exams. Attaching the test to this procedure helped ensure the students would take everything seriously but also introduced a few complications. The EI test began with some sample items and instructions for calibrating the headset, which allowed students to become familiar with the task type and troubleshoot issues with the equipment. Despite this pre-test calibration, five students still experienced technical difficulties that made their responses impossible to score. These students were excluded from the final analysis. Due to the large number of students who completed the task simultaneously—up to 54 students at a time—participants were likely to hear a certain degree of background noise in spite of their headsets; however, all participants experienced these same conditions so there is little evidence that the noise affected responses systematically. The large number of simultaneous tests being administered introduced another difficulty, however, as it was possible for students to bypass the pre-test survey and proctors were unable to intervene. Thus, only a subset of 136 of the possible 213 students completed the survey.

Test scoring

The tests were all single-rated by one researcher, and a team of 12 other trained raters provided second ratings. Training included a detailed explanation of the task followed by a supervised rating of 15 practice items. The scoring was done digitally so that raters could simultaneously listen to and score each sentence uttered by a student. When raters listened to the recordings, they attempted to match what they heard with the text of the prompt as they viewed it on a screen. The result was that small errors in pronunciation did not count against test-takers as long as what they said was identifiable. For example, if a test-taker inserted a vowel into the consonant cluster in the word *store* (possibly resulting in the word being pronounced /sətoʊɹ/), the test-taker would have received credit for correctly uttering the single-syllable word *store*.

The database program then reported the average of the percentage of syllables correctly repeated, from the raters without regard for the order of the utterance, and it was subsequently converted to a 4-point Andrich rating scale with the levels based on the number of correct syllables with 0 = None (0% correct), 1 = Up to half (1% to 49%), 2 = More than half (50% to 99%), and 3 = All (100%). For example, the hypothetical sentence, "I was walking to the store yesterday," includes 10 syllables; if a student heard this sentence and then repeated "I walked to the store yesterday" the student would score 80% since two of the 10 syllables were missing (e.g., *was* and *-ing*) and would have the category of 2. Also, the lack of ordering constraints allowed students to repair an erroneous repetition; for example, if a student repeated, "I went to the store yesterday—I was walking," then full points would be awarded.

Data analysis

To answer the first research question, the results of the EI test were analyzed using FACETS (Linacre, 2012) with four facets: Examinee, Item, Accent and Intended Item Level (a dummy facet used since items are nested within the intended item level). Dummy facets do not contribute to the calculation of measurement; however, they allow researchers to investigate interactions. An ANOVA was run with the Intended Item Level facet as grouping factor to allow us to see if those items functioned as hypothesized (e.g., Intermediate easier than Advanced, and both easier than Superior). In addition to examining the interaction between the accent and an item's intended difficulty level, a Rasch bias analysis was conducted. To answer the second research question, the results of the pre-test survey, along with the Accent measures, were analyzed using descriptive statistics, one-way ANOVAs, and a Spearman's Rho correlation.

Results

Before addressing the research questions, it is important to determine how well the test functioned. Rasch analysis was used to evaluate the instrument including an analysis of the rating scale, an evaluation of the fit statistics, and the reliability of the measurements in the different facets.

The FACETS program produces a display of data known as a *Wright Map* (see Figure 4) that places all the facets on a single, comparable scale. The first column on the vertical scale indicates the logit measurement, which is based on the mean performance of the examinees (the mean is indicated by 0). The second column represents the Examinee facet and shows the variability in examinee scores (each * indicating 3 examinees). The third column shows Accent; the fourth column, Items; and the fifth column is the dummy facet that shows Intended Item Level. The sixth column maps the 4-Point Andrich EI Rating scale to the logits.

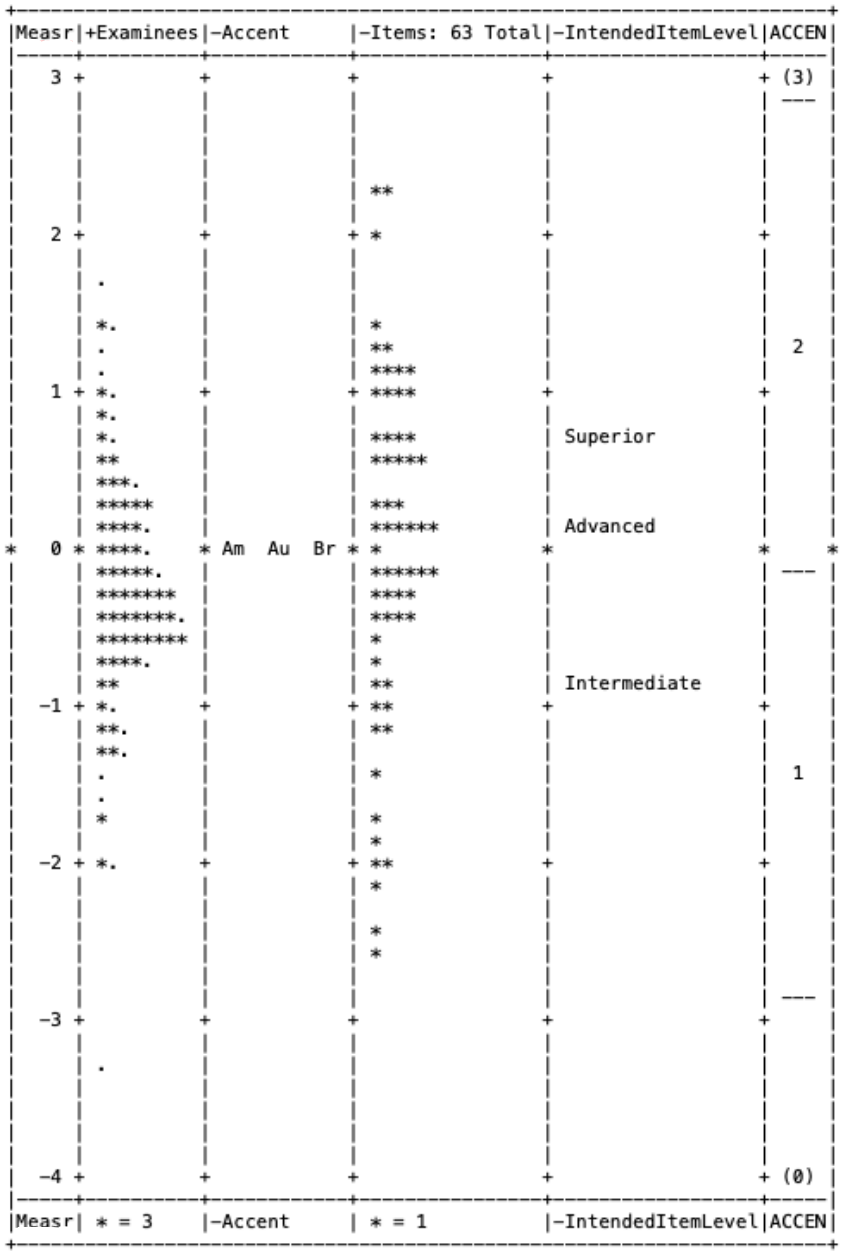

Figure 4. Wright map of examinees, accent, EI Items and intended item level

Notes: Columns: (1) logit, (2) examinee performance, (3) accent, (4) item, (5) intended item level and (6) Andrich EI Rating Scale.

Accent key: Am=American, Au=Australian, Br=British

Based on the rating scale diagnostic guidelines in McNamara et al. (2019), the 4-point rating scale functioned adequately in that it had (1) enough observations at each level, (2) average measures that advanced at approximately the same values at each score point, (3) means that were not disordered, (4) frequency data points that resulted in a smooth distribution (see Figure 5), (5) average measures near their expected values, (6) outfit mean square values less than 2.0, and (7) Rasch-Andrich thresholds that increased by at

least 1.4 logits but no more than 5 logits (see Table 3).

Table 3. Rating scale diagnostic for 4-point Andrich EI rating scale

Label	Score	Observed Count	Perc	Obsvd Avrg	Sample Expect	Outfit MNSQ	Andrich Threshold	Category Measure
None	0	2920	14%	-1.93	-2.17	1.3	None	-3.72
Up to half	1	8314	40%	-1.10	-0.94	0.9	-2.60	-1.38
More than half	2	7560	36%	0.61	0.54	0.9	-0.15	1.31
All	3	1975	10%	2.24	2.21	1.0	2.75	3.87

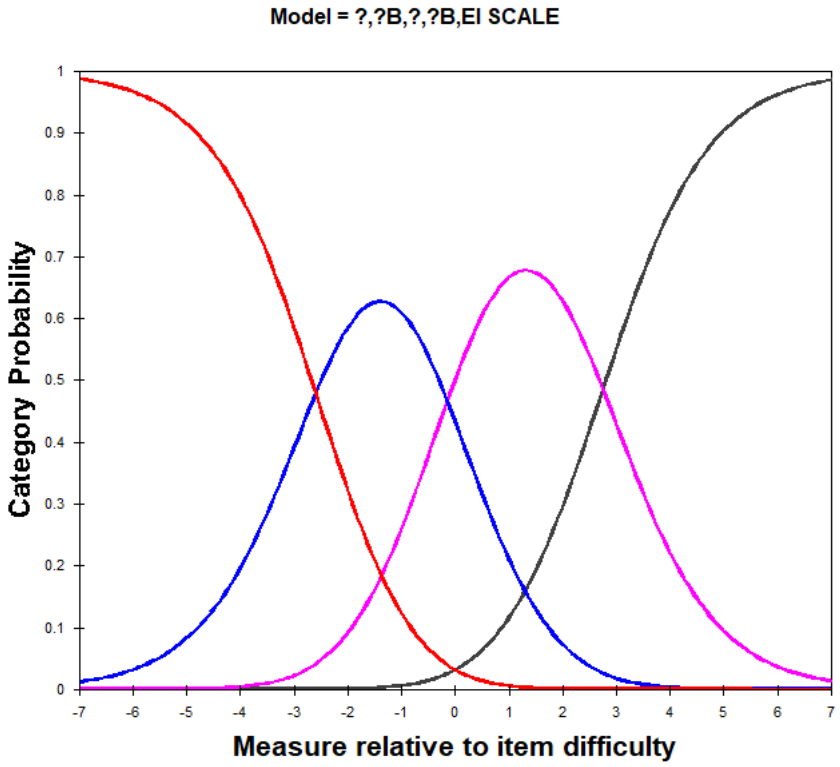

Figure 5. Probability curves for 4-point EI rating scale

To examine the fit of the data to the model, the fit statistics of the four facets were analyzed. For three of the facets, Item (n = 63), Accent (n = 3) and Intended Item Levels (n = 3), there were no instances of overfit (i.e. Outfit or Infit Mean Squares (MNSQ) < .5) or underfit (i.e. Outfit or Infit Mean Squares (MNSQ) > 1.5). With the 213 examinees, there were 8 examinees (3.7%) that overfit the model and 27 examinees (12.7%) that underfit the model. To determine if those misfitting persons impacted measurement (Linacre, 2010), the misfits were excluded from the analysis and the person measures were estimated again and subsequently cross-plotted with the measures obtained using all the data to see if there were any noticeable changes. After removing the misfitting persons, this cross-plotting of measures was performed on the remaining 178 examinees, with a

correlation coefficient of 1.0. This indicates that the misfitting data did not negatively impact measurement; subsequently the entire dataset, including those examinees that did not fit, was used.

The initial MFRM analysis found that, for the facets used to calculate measurement, (Examinee, Item, and Accent), the subsets were connected. The program sets the means of each facet to 0 except the one that is allowed to float, which in most instances is the Examinee facet. Since the mean of the Examinee facet was -.23, the items were slightly more difficult than the examinee ability. Fixed model chi-square tests indicate that all of the elements in these facets have statistically different measures (see Table 4). For separation reliability, the closer the value approaches 1.0, the greater likelihood that that differences between the elements in a facet are due to varying trait levels. In essence, these reliability coefficients are analogous to interpreting *r* where r^2 indicates effect size. Thus, the square of the reliability coefficient can be used to explain the variance. The Item facet was 1.00 with a Separation Ratio of 19.72, indicating there would be over 19 distinct levels of item difficulty. Since there were only 3 intended item difficulty levels, the value seems quite high, but the formula used to calculate this statistic divides the True SD by the Root Mean-Square Error which was quite low (.10). Thus on a practical level, these high values indicate that the criteria used to create at least three distinct item difficulty levels functioned as anticipated. The Examinee facet was .94 indicating that 88.4% of the variance in test scores can be attributed to differences in examinee ability.

Table 4. Separation reliability statistics for examinees, items & accent

	Examinees	Items	Accent
	N = 213	N = 63	N = 3
Measures			
Mean	-.23	.00	.00
SD	.73	1.52	.04
Infit			
Mean	1.01	.99	1.00
SD	.46	.20	.01
Outfit			
Mean	1.00	.99	1.00
SD	.45	.20	.01
Number with Outfit > 1.5	27	0	0
Separation statistics			
Separation Reliability	.94	1.00	.79
Separation Strata	5.83	19.72	2.96
SE Mean (RMSE)	.17	.10	.02
Chi-square	3,594.10	15,489.80	14.60
df	212	62	2
p	.00	.00	.00

Research question 1

The first research question examined the extent to which accent impacted item difficulty. With a separation reliability of .79, the Accent Facet did result in statistically different measures (see Table 4). Table 5 presents the Accent facet measurement report and shows that the American accent was found to be the easiest (-0.06) and the British accent the most difficult (0.05). The dummy facet of Intended Item Levels also had reliably different measures with the Intermediate items being the easiest (-0.83) and Superior items the most difficult (0.67).

Table 5. Facet measurement report for accent and intended item level

	Total Count	Obs. Avge	FairM Avge	CTT Item Difficulty	Measure	Model S.E.	Infit MnSq	Outfit MnSq
Accent								
American (AM)	6919	1.43	1.47	0.48	-0.06	0.02	1.00	1.00
Australian (Au)	6926	1.42	1.44	0.47	0.01	0.02	.99	.99
British (Br)	6924	1.39	1.43	0.46	0.05	0.02	1.01	1.01
Mean	6923	1.41	1.45	0.47	0.00	0.02	1.00	1.00
SD	2.9	0.02	0.02	0.01	0.04	0.00	.01	.01
Intended Item Level								
Intermediate (Int)	6979	2.01	1.75	0.67	-0.83	0.02	1.10	1.08
Advanced (Adv)	6931	1.31	1.39	0.44	0.15	0.02	1.01	1.01
Superior (Sup)	6859	0.91	1.19	0.30	0.67	0.02	.89	.91
Mean	6923	1.41	1.44	0.47	0.00	0.02	1.00	1.00
SD	49.3	0.46	0.23	0.15	0.62	0.00	.08	.07

An ANOVA found the main effect of Intended Item Level was statistically significant [$F(2, 20{,}760) = 17{,}176.81$, $p < .001$] as was Accent [$F(2, 20{,}760) = 22.98$, $p < .001$]. (See Figure 6). For Intended Item Level, a Tukey Post Hoc test found significant differences with large effect sizes between Intermediate and Advanced [mean difference = 2.00, $t = 103.62$, $p < .001$, *Cohen's D* = 1.76] and Advanced and Superior [mean difference = 1.11, $t = 57.10$, $p < .001$, *Cohen's D* = .97]. A Tukey Post Hoc test only found significant differences in Accent between American and Australian [mean difference = 0.08, $t = 3.97$, $p < .001$, *Cohen's D* = .07] and American and British [mean difference = 0.11, $t = 5.83$, $p < .001$, *Cohen's D* = .010], though the effect sizes were small.

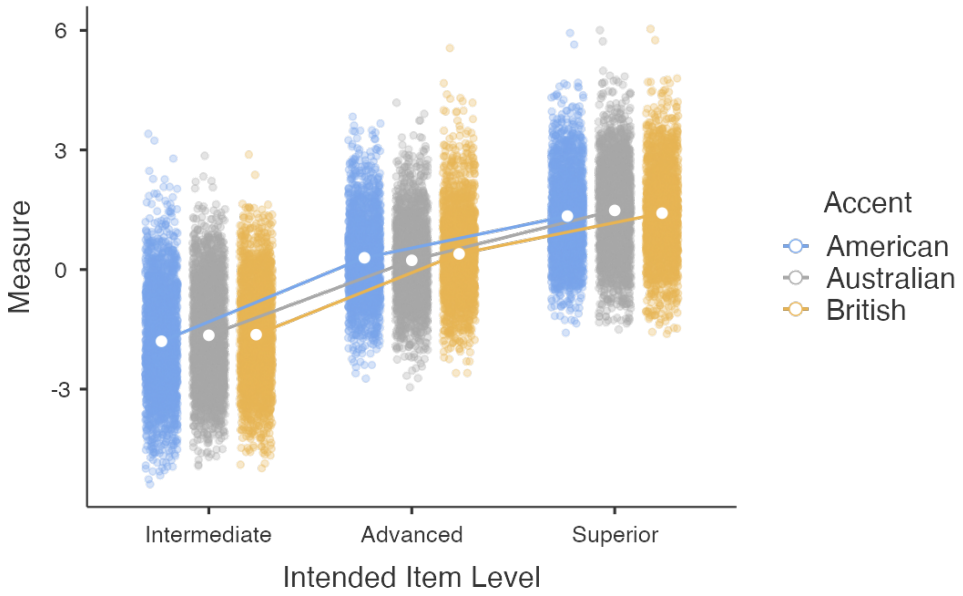

Figure 6. Marginal means of intended item level and accent

To see if there was an interaction between Accent and Intended Item Level, a bias/interaction analysis was conducted. In Table 6, the t-statistic is reported in the 7th column along with the probability value in the 9th column. We see that there are no significant interactions between the accent and intended item level.

Table 6. Bias/Interaction Report of Accent and Intended Item Level (arranged by Bias Size)

		Observd Score	Expctd Score	Observd Count	Obs-Exp Ave	Bias+ Size	Model S.E.	*t*	d.f.	*p- value*	Infit MnSq	Outfit MnSq
Aus	Sup	2018	1999.60	2258	0.01	0.02	0.04	0.66	2257	0.51	1.0	1.0
Am	Adv	3018	2998.62	2280	0.01	0.02	0.03	0.67	2279	0.51	1.0	1.0
Br	Int	4527	4513.97	2266	0.01	0.02	0.04	0.48	2265	0.63	1.1	1.1
Au	Int	4824	4829.26	2417	0	-0.01	0.04	-0.19	2416	0.85	1.0	1.0
Br	Adv	3067	3073.33	2400	0	-0.01	0.03	-0.21	2399	0.83	1.0	1.0
Br	Sup	2045	2051.92	2258	0	-0.01	0.04	-0.25	2257	0.81	0.9	0.9
Am	Int	4685	4692.61	2296	0	-0.01	0.04	-0.28	2295	0.78	1.2	1.1
Am	Sup	2176	2187.91	2343	-0.01	-0.01	0.03	-0.41	2342	0.68	0.8	0.9
Au	Adv	2999	3012.34	2251	-0.01	-0.02	0.03	-0.46	2250	0.65	1.0	1.0
	Mean	3262.1	3262.17	2307.7	0.00	0.00	0.04	0.00			1.0	1.0
	S.D.	1077	1077.51	60.1	0.01	0.02	0.00	0.44			0.1	0.1

Research question 2

To answer the second research question, regarding the effect of familiarity on EI test performance, the pre-test survey (n = 136) results were analyzed and students were

divided into bands of familiarity. Figure 7 presents violin plots[1] of the data and shows that students were most familiar with American English, least familiar with Australian and more equally distributed with British English. This is unsurprising as 2/3rds of the students were from the Americas.

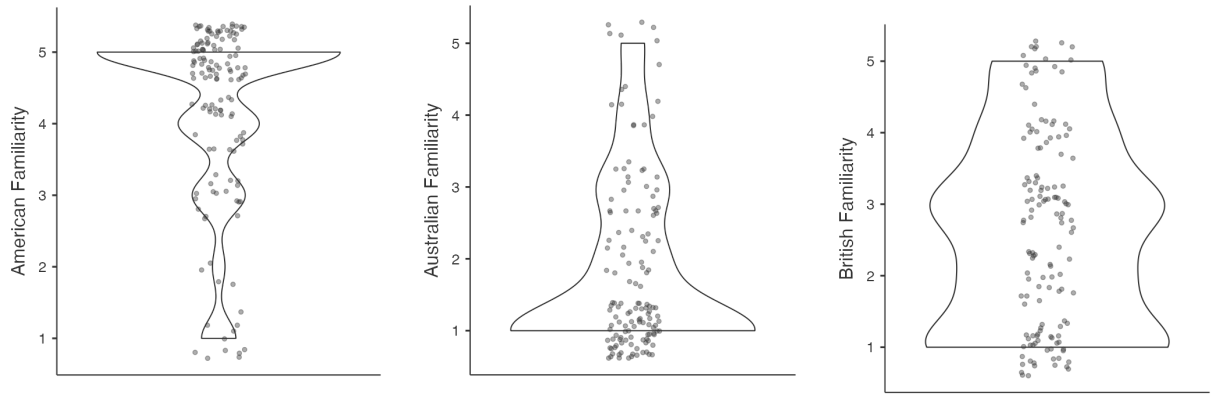

Figure 7. Results of pre-test survey (n=133)
Notes: Responses are to the question "Overall, how familiar are you with the following English accents?" that were anchored with 1 = "Not at all familiar"; 5 = "Familiar." Note that no verbal descriptors were provided to the examinees for options 2, 3 or 4.

An examination of the EI observed average scores on the subtests for the three varieties found that the more familiar examinees were with a particular accent, the more likely they were to have a higher score on items with those accents (see Table 7). This was particularly true for those who were *Familiar* (Likert Scale category 5) compared to those who were *Not at all Familiar* (Category 1).

Table 7. Descriptive statistics of EI accent measures by accent familiarity

Accent Familiarity	EI Measure American English				EI Measure Australian English				EI Measure British English			
	N	Mean	SE	SD	N	Mean	SE	SD	N	Mean	SE	SD
1—Not at All Familiar	11	-0.73	0.39	1.29	77	-0.26	0.08	0.70	39	-0.41	0.13	0.79
2	4	-0.33	0.20	0.40	21	-0.01	0.13	0.60	24	-0.16	0.13	0.66
3	17	-0.45	0.14	0.58	22	-0.04	0.19	0.90	36	-0.10	0.12	0.72
4	26	-0.30	0.14	0.71	9	-0.07	0.22	0.66	20	0.12	0.16	0.72
5—Familiar	78	0.04	0.10	0.84	7	0.15	0.23	0.60	17	-0.08	0.23	0.94

To determine if this finding was statistically significant, three one-way ANOVAs were conducted between the IV accent familiarity and the DV EI Accent measures. The participants that were more familiar with American English had a significant difference

[1] Similar to boxplots except they also show the probability density.

with the American EI Accent Measure [F (4, 131) = 3.16, p = .016]; however, there was not a significant difference with Australian English and Australian EI Accent Measure [F (4, 131) = 1.15, p = .335], nor British English with British EI Accent Measure [F (4, 131) = 1.80, p = 0.132].

Since level of familiarity might not represent distinct categories of people, the more conservative Spearman's Rho correlation was also run to cross validate if the relationship between familiarity and test score was significant. The relationships between the familiarity with each EI accent with their respective EI subtest scores were with the American (*Spearman's Rho* = .27, p < .001), Australian (*Spearman's Rho* = .20, p = .02), and British (*Spearman's Rho* = .22, p = .012), indicating a moderate positive correlation. In every instance, accent familiarity seems to have a positive relationship with higher person ability estimates on the EI items that used that accent. Note that examinees were not asked to rank their familiarity of accents so that it is possible that some examinees indicated high levels of familiarity for all three accents and others could have indicated the inverse. It is possible that the more proficient an examinee is, the more familiar they would be with multiple varieties while the less proficient examinees would likely be less familiar with any variety.

Discussion and conclusion

This research confirms what other studies have found: unfamiliar accents hinder listening comprehension (Harding, 2018; Ockey et al., 2016, Ockey & Wagner, 2018) and this effect extends to EI item types as well. The first research question investigated the effect of accent on EI test difficulty. The results found that accent familiarity might impair or facilitate performance. For the group of students in this study, British English was the most difficult, Australian English was the second most difficult, and American English was the easiest to understand. Furthermore, this was found across different intended item difficulty levels. Items designed to be easier (Intermediate) or more difficult (Superior) were similarly affected by the accent even though the effect size was small. The second research question investigated how students' accent familiarity affected their performance with the different accents. Students who were unfamiliar with an accent tended to perform more poorly on items in that accent. Test developers, therefore, need to take accent into account as a factor that can systematically affect the item difficulty of any given EI prompt.

Accounting for accent is of particular importance as EI testing emerges as an affordable, low-stakes general proficiency assessment. What does the test claim to measure? Does it measure English as a *lingua franca* or English for a specific context? As the item difficulty varies based on examinee familiarity to accent, an exam with only one accent might

systematically favor or harm the examinees by underrepresenting the construct. Including multiple accents in an EI test will likely raise other challenges for test creators—like how to choose the exact selection of accents that is fair for all test takers (Elder & Harding, 2008)—but for those to continue to use single-accent tests it may be beneficial to provide justification for the increase in difficulty to students unfamiliar with that accent. If an EI test is administered for a specific institution in which one variety of English is the norm, it could be argued that the test is designed to assess language ability in a local context. However, claims of EI that contain a single accent to measure global proficiency need to be carefully scrutinized.

For ease of test development, audio prompts are often transcribed and test developers might not regularly listen to the actual prompts and note their acoustic qualities. In such instances, developers run the risk of assuming the item difficulty parameters are immutable based solely on the linguistic characteristics of the written text (e.g., vocabulary, number of syllables, syntactic complexity) without considering the transformation that occurs when the modality changes to an oral text. Along with accent, other issues to be accounted for could include rate of speech, reductions, enunciation, etc. Assuming that the item difficulty parameter would be the same based solely on a written representation is problematic and may be happening unawares, especially if one thinks that a script could be used for onsite testing that a proctor might read. Thus, item difficulty parameters need to be tied to the actual audio presented to the examinee and not the written representation of the prompt.

This study also has implications for ESL assessment with listening components generally. Previous studies have already demonstrated that accent can affect listening comprehension in a TOEFL setting (e.g., Ockey & French, 2014; Ockey et al., 2016)—that is, with tasks that require participants to listen to passages and answer multiple-choice questions about the content—but the results of the current study indicate that accent affects comprehension in different types of listening assessments as well.

There are some limitations to this study that should be considered. First, Derwing and Munro's (2009) definition of *accent* views accents as features of the observer's perception rather than features of speech. In this study, native speakers were used to determine strength of accent, and it was assumed that this represented how the speech samples were perceived by nonnative speakers from a wide range of proficiency and L1 backgrounds. No studies have yet confirmed that listeners perceive an L2 accent in the same way that they perceive an L1 accent. Second, rate of speech was not accounted for in the creation of the EI prompts. Previous research has demonstrated that rate of speech can affect listening comprehension (Anderson-Hsieh & Koehler, 1988), and it was not controlled for in the present study.

The findings of this study confirm previous research that suggests that accent can impair listening comprehension (e.g., Anderson-Hsieh & Koehler, 1988; Floccia et al., 2006; Varonis & Gass, 1982). It also adds to the very limited body of research on the effects of regional or international accents in listening in a second language (Ockey & French, 2014; Major et al., 2005) and adds another dimension by showing there is an effect in the context of EI testing. Since EI tests require examinees to not only understand but produce oral responses, the role of accent likely affects other listening item types as well. However, since the effect size was small on the item difficulty parameters, the impact of accent on individual test scores might affect lower level examinees more than those with higher proficiency. Test developers should consider including accent in the test and item specifications of their listening measurement instruments.

References

Abeywickrama, P. (2013). Why not non-native varieties of English as listening comprehension test input? *RELC Journal, 44*(1), 59–74. http://dx.doi.org/10.1177/0033688212473270

Adank, P., & McQueen, J. M. (2007). *The effect of an unfamiliar regional accent on spoken word comprehension.* Proceedings of the 16th International Congress of Phonetic Sciences. International Congress of Phonetic Sciences: Saarbrücken, Germany. http://www.icphs2007.de/conference/Papers/1387/1387.pdf

Adank, P., Evans, B. G., Stuart-Smith, J., & Scott, S. K. (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *Journal of Experimental Psychology: Human Perception and Performance, 35*(2), 520. http://dx.doi.org/10.1037/a0013552

Anderson-Hsieh, J., & Koehler, K. (1988). The effect of foreign accent and speaking rate on native speaker comprehension. *Language Learning, 38*(4), 561–613. http://dx.doi.org/10.1111/j.1467-1770.1988.tb00167.x

Brunfaut, T., & Révész, A. (2015). The role of task and listener characteristics in second language listening. *TESOL Quarterly, 49*(1), 141–168. https://doi.org/10.1002/tesq.168

Chan, D., Schmitt, N., DeShon, R. P., Clause, C. S., & Delbridge, K. (1997). Reactions to cognitive ability tests: The relationships between race, test performance, face validity perceptions, and test-taking motivation. *Journal of Applied Psychology, 82*(2), 300. http://dx.doi.org/10.1037/0021-9010.82.2.300

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America, 116*(6), 3647–3658. http://dx.doi.org/10.1121/1.1815131

Cox, T., Bown, J., & Burdis, J., (2015). Exploring proficiency-based versus language for specific purposes items with elicited imitation assessment. Foreign Language Annals, 48(3), 350-371. https://doi.org/10.1111/flan.12152

Cox, T., & Davies, R. (2012). Using automatic speech recognition technology with elicited oral response testing. *CALICO Journal, 29*(4), 601-618. https://www.jstor .org/stable/calicojournal.29.4.601

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility. *Studies in Second Language Acquisition, 19*(1), 1–16. http://dx.doi.org/10.1017 /S0272263197001010

Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching, 42*(4), 476–490. http://dx.doi.org/10.1017 /S026144480800551X

Elder, C., & Harding, L. (2008). Language testing and English as an international language: Constraints and contributions. *Australian Review of Applied Linguistics, 31*(3), 1-11. http://dx.doi.org/10.2104/aral0834

Floccia, C., Goslin, J., Girard, F., & Konopczynski, G. (2006). Does a regional accent perturb speech processing? *Journal of Experimental Psychology: Human Perception and Performance, 32*(5), 1276-93. http://dx.doi.org/10.1037/0096 -1523.32.5.1276

Gass, S., & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning, 34*(1), 65–87. http://dx.doi.org/10.1111/j .1467-1770.1984.tb00996.x

Graham, C. R., Lonsdale, D., Kennington, C. R., Johnson, A., & McGhee, J. (2008, May). *Elicited Imitation as an oral proficiency measure with ASR scoring.* Paper presented at the Sixth Language Resources and Evaluation Conference (LREC), Marrakech, Morocco. http://www.lrec-onf.org/proceedings/lrec2008/pdf/409 _paper.pdf

Harding, L. (2011). *Accent and listening assessment: A validation study of the use of speakers with L2 accents in academic English listening assessment*. Peter Lang. https://eprints.lancs.ac.uk/id/eprint/50928/

Harding, L. (2018). Listening to an unfamiliar accent: Exploring difficulty, strategy use, and evidence of adaptation on listening assessment tasks. In G. J. Ockey & E. Wagner (Eds.), *Assessing L2 listening: Moving towards authenticity.* John Benjamins. http://dx.doi.org/10.1075/lllt.50.07har

Kang, O., Thomson, R., & Moran, M. (2019). The effects of international accents and shared first language on listening comprehension tests. *TESOL Quarterly, 53*(1), 56-81. http://dx.doi.org/10.1002/tesq.463

Linacre, J. M. (2012). *Winsteps® (Version 3.75. 0) [Computer Software].* Beaverton, Oregon: Winsteps.com. Retrieved May 12, 2021. Winsteps.com

Linacre, J. M. (2010). When to stop removing items and persons in Rasch misfit analysis?, *Rasch Measurement Transactions, 23*(4), 1241. https://www.rasch.org /rmt/rmt234g.htm

McNamara, T., Knoch, U., & Fan, J. (2019). *Fairness, Justice, and Language Assessment.* Oxford University Press.

Major, R. C., Fitzmaurice, S. F., Bunta, F., & Balasubramanian, C. (2002). The effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL Quarterly, 36*(2), 173–190. http://dx.doi.org/10.2307/3588329

Major, R. C., Fitzmaurice, S. M., Bunta, F., & Balasubramanian, C. (2005). Testing the effects of regional, ethnic, and international dialects of English on listening comprehension. *Language Learning, 55*(1), 37–69. http://dx.doi.org/10.1111/j.0023 -8333.2005.00289.x

Matsushita, H., & Lonsdale, D. (2012). *Item development and scoring for Japanese oral proficiency testing.* Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC '12), European Language Resources Association (ELRA), 2682–2689.

Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008). The weckud wetch of the wast: Lexical adaptation to a novel accent. *Cognitive Science, 32*(3), 543–562. http://dx .doi.org/10.1080/03640210802035357

Millard, B., & Lonsdale, D. (2014). French oral proficiency assessment. In *Variation within and across romance languages: Selected papers from the 41st linguistic symposium on romance languages (LSRL)*, Ottawa, 5–7 May 2011 (Vol. 333, pp. 401). John Benjamins Publishing Company. http://dx .doi.org/10.1075/cilt.333.26mil

Moulton, S. E. (2012). *Elicited Imitation testing as a measure of oral language proficiency at the Missionary Training Center.* Theses and Dissertations. Paper 3137. http://scholarsarchive.byu.edu/etd/3137

Mugglestone, L. (2007). *Talking proper: The rise of accent as social symbol*. Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780199250622 .001.0001

Munro, M. J., & Derwing, T. M. (1995a). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 45*(1), 73–97. http://dx.doi.org/10.1111/j.1467-1770.1995.tb00963.x

Munro, M. J., & Derwing, T. M. (1995b). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech, 38*(3), 289–306. http://dx.doi.org/10.1177/002383099503800305

Ockey, G. J. (2018). Reliability and sources of score variance in a strength of accent measure. In G. J. Ockey, & E. Wagner (Eds.), *Assessing L2 listening: Moving towards authenticity*. John Benjamins. http://dx.doi.org/10.1075/lllt.50.06ock

Ockey, G. J., & Wagner, E. (2018). An overview of the issue of using different types of speech varieties as listening inputs in L2 listening assessment. In G. J. Ockey & E. Wagner (Eds.), *Assessing L2 listening: Moving towards authenticity.* John Benjamins. https://benjamins.com/catalog/lllt.50.c5

Ockey, G. J., & French, R. (2014). From one to multiple accents on a test of L2 listening comprehension. *Applied Linguistics, 37*(5), 693-715. http://dx.doi.org/10.1093/applin/amu060

Ockey, G. J., Papageorgiou, S., & French, R. (2016). Effects of Strength of Accent on an L2 Interactive Lecture Listening Comprehension Test. *International Journal of Listening, 30*(1-2), 84-98. http://dx.doi.org/10.1080/10904018.2015.1056877

Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing, 29*(3), 325-344. http://dx.doi.org/10.1177/0265532211424478

Varonis, E. M., & Gass, S. (1982). The comprehensibility of non-native speech. *Studies in Second Language Acquisition, 4*(02), 114–136. http://dx.doi.org/10.1017/S027226310000437X

Vinther, T. (2002). Elicited Imitation: A brief overview. *International Journal of Applied Linguistics, 12(*1), 54–73. http://dx.doi.org/10.1111/1473-4192.00024

Wingstedt, M., & Schulman, R. (1984). Comprehension of foreign accents. *Phonologica,* 339–345.

Yan, X., Maeda, Y., Lv, J., & Ginther, A. (2016). Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. *Language Testing, 33*(4), 497-528. http://dx.doi.org/10.1177/0265532215594643

Appendix

Q1 Overall, how familiar are you with these English accents?

	Not at all familiar (1)	(2)	(3)	(4)	Familiar (5)
American (US) (1)	○	○	○	○	○
Australian (2)	○	○	○	○	○
British (UK) (3)	○	○	○	○	○
Other native accents (e.g., Canadian, New Zealander, etc.) (4)	○	○	○	○	○
Non-native accents (5)	○	○	○	○	○

Q2 How often have you heard these English accents on TV, radio, the internet, or other media?

	Rarely (1)	(2)	(3)	(4)	Very often (5)
American (US) (1)	○	○	○	○	○
Australian (2)	○	○	○	○	○
British (UK) (3)	○	○	○	○	○
Other native accents (e.g., Canadian, New Zealander, etc.) (4)	○	○	○	○	○
Non-native accents (5)	○	○	○	○	○

Q3 How often have you heard these English accents in face-to-face communication?

	Rarely (1)	(2)	(3)	(4)	Very often (5)
American (US) (1)	○	○	○	○	○
Australian (2)	○	○	○	○	○
British (UK) (3)	○	○	○	○	○
Other native accents (e.g., Canadian, New Zealander, etc.) (4)	○	○	○	○	○
Non-native accents (5)	○	○	○	○	○

Q4 How long have you studied English with teachers who have these accents?

	Not at all (1)	Less than 1 year (2)	1-2 years (3)	3-4 years (4)	5+ years (5)
American (US) (1)	○	○	○	○	○
Australian (2)	○	○	○	○	○
British (UK) (3)	○	○	○	○	○
Other native accents (e.g., Canadian, New Zealander, etc.) (4)	○	○	○	○	○
Non-native accents (5)	○	○	○	○	○

Q5 How long have you lived in these English-speaking countries?

	Not at all (1)	Less than 6 monts (2)	6-12 months (3)	1-2 years (4)	3+ years (5)
United States (1)	○	○	○	○	○
Canada (2)	○	○	○	○	○
Australia (3)	○	○	○	○	○
United Kingdom (4)	○	○	○	○	○
Other (please be specific) (5)	○	○	○	○	○