tesol QUARTERLY

The Effects of International Accents and Shared First Language on Listening Comprehension Tests

OKIM KANG AND MEGHAN MORAN
Northern Arizona University
Flagstaff, Arizona,United States

RON THOMSON
Brock University
St. Catharines, Ontario, Canada

This study examines the effect of incorporating a variety of international English accents into a simulated TOEFL listening comprehension test in growing recognition of internationalization of language teaching and learning in the field of TESOL. Although some high-stakes English proficiency exams have begun incorporating speech samples produced by speakers from a range of inner circle English-speaking backgrounds (e.g., Britain, the United States, Australia), the inclusion of samples produced by speakers of outer and expanding circle English varieties (e.g., India, Nigeria, Mexico, South Korea) has been largely avoided. For this study the researchers recruited speakers from six distinct English varieties to produce speech samples for a mock TOEFL iBT listening exam. Listeners who spoke with the same six international English accents were then recruited to take the resulting tests. Results indicate that when accented English is highly comprehensible, listening test scores for stimuli based on high-proficiency speakers of outer and expanding circle varieties of English are not significantly lower than they are in response to stimuli based on inner circle varieties of English. With respect to a shared first language effect on test scores when test materials are spoken in the test taker's own accent, results are complex but inconclusive.

doi: 10.1002/tesq.463

International tests of English proficiency have been criticized on the grounds that they privilege a putative "standard" variety of English (e.g., Received Pronunciation [RP], General American [GA]) and are therefore unfair to test takers who speak a nonstandard English variety (e.g., Indian, Nigerian; Hamp-Lyons & Davies, 2008). Given a growing recognition that many valid and widely used varieties of English have emerged, Jenkins (2006) goes so far as to argue that there is no longer a rationale for speakers from nontraditional English-speaking contexts, in which English language teaching is a developing field, to defer to native speakers and their particular standards of English. Following similar reasoning, scholars such as Taylor (2006) have suggested that English proficiency tests should now adopt an English as an international language (EIL) approach over reference to traditionally standard varieties (Hamp-Lyons & Davies, 2008). Specifically, English language assessment conventions need to adapt in order to recognize EIL's greater heterogeneity of norms.

Especially relevant to this new movement toward English as an international language and the growing acceptance of World Englishes (see Kachru, 1992) is the assessment of listening skills. Given that foreign-born instructors are commonplace in North American colleges and universities (Fitch & Morgan, 2003) and elsewhere, an ecologically valid test of English listening would necessarily require that listeners be able to understand academic lectures spoken in a variety of English accents. Therefore, only relying on North American–accented English speech in the creation of the Test of English as a Foreign Language (TOEFL) listening stimuli could underrepresent the variety of English accents found in the target domain. To the extent that underrepresented accents impact task performance, this could weaken task validity in predicting real-world performance; that is, reliance on traditionally accepted English norms overlooks the sociolinguistic realities of candidates' ultimate language use (Elder & Harding, 2008).

With the rise of English as a localized language in many international contexts, speakers around the world have adapted the language for their own uses, creating distinct varieties of English. In this study, we focus solely on the phonological aspects of international English varieties (i.e., accents) to the exclusion of other dialect features, such as differences in syntax, morphology, the lexicon, and discourse. This focus is deliberate and a consequence of the fact that listening passages in English language tests are typically scripted according to a standard English dialect, but use a variety of voices. Furthermore, whereas listener comprehension of differences in syntax, morphology, vocabulary, and so on are more learnable, phonological differences are known to be far more resistant to change in adult learners (Hyltenstam & Abrahamsson, 2000), meaning that a bias toward the use of a particular accent on language tests may disadvantage learners who speak—and therefore comprehend through—a different accent.

Given the challenge of acquiring a second language (L2) sound system, more needs to be understood concerning the extent to which listeners from one first language (L1) background respond to speech spoken with the same accent vis-à-vis different accents. Several studies have reported a shared L1 benefit, suggesting that processing speech spoken in one's own accent is easier than processing speech spoken with a different accent (e.g., Bent & Bradlow, 2003; Kang, 2012; Munro, Derwing, & Morton, 2006). Because these previous studies were conducted under laboratory conditions, the degree to which a shared L1 effect plays a role in the context of real-world listening tests remains uncertain. Furthermore, we are unaware of previous studies having investigated whether there is an interaction between this shared L1 benefit and measures of the speech samples' comprehensibility and strength of accentedness. That is, does the strength of the shared L1 benefit weaken if the accented speech is easy to process by listeners from other L1s and is less divergent, phonologically, from American English? Or do such shared L1 effects still exist if speakers are highly comprehensible in their speech? We attempt to find answers to these queries by empirically examining interactions between speech with varied accents by listeners from varied L1 backgrounds in listening comprehension tests.

The current study explores the impact of L1 accent on listeners' comprehension scores in an authentic high-stakes test, in which speech samples spoken with a variety of English accents are incorporated. The study's aim is to inform the design of future content for high-stakes listening tests. This approach will build connections between score interpretations and real-world behaviors. More specifically, we examine the impact of different varieties of English pronunciations on listeners' comprehension scores on the TOEFL internet-based test (iBT) listening section and interactions between English varieties used as test samples, the L1 origin of test takers, and the comprehensibility/strength of accent of the speech they produce.

VARIETIES OF ENGLISH IN ASSESSMENT OF LISTENING COMPREHENSION

In the listening assessment literature, the role of the test speaker's L1 accent has been seen as a multifaceted area of inquiry for more than a decade (Llurda, 2004). On one hand, there has been a push for the inclusion of L1 accent varieties in the listening sections of high-stakes English exams (Abeywickrama, 2013; Harding, 2012; Ockey & French, 2016; Ockey, Papageorgiou, & French, 2016). With approximately 505 million nonnative English speakers (OMICS International, 2013), many native English speakers (NESs) as well as nonnative

English speakers (NNESs) are exposed to a wide range of accents in their everyday interactions. Therefore, the inclusion of a variety of English accents in listening stimuli is warranted on "the bases of enhanced authenticity [of the exams], a more accurate representation of the listening construct, and the potential for positive washback" (Harding, 2012, p. 164).

However, researchers have noted the complexities and potential drawbacks of including international English accents as well. For instance, the possibility of test bias as well as logistical concerns (e.g., inappropriate sample) and random errors have caused test developers to be more conservative in their approach (Ockey & French, 2016; Taylor, 2006; Taylor & Geranpayeh, 2011) to selecting a variety of accents in the target language use domain. To achieve a fair, ecologically valid English assessment, a balance must be struck between domain representation and reliability.

That being said, some global tests of English listening skills have begun experimenting with the inclusion of non-RP and non-GA test items in recent years (e.g., IELTS, TOEIC). Although we acknowledge the strides these tests are taking to be more inclusive and authentic, their attempts are problematic in that they still only incorporate inner circle English accents. That is, in the three most prominent English proficiency exams (i.e., TOEFL iBT, IELTS, and TOEIC), even the newly incorporated accents are considered prestigious varieties, despite the tests' stated desires to reflect authentic English communication. These tests assert that their inclusion of Canadian, Australian, and New Zealand speakers increases the tests' authenticity in this increasingly globalized world, but it seems unlikely that these few varieties are representative of the majority of English communication in international contexts.

THE EFFECT OF ACCENT ON LISTENING COMPREHENSION

Recent research in applied linguistics and TESOL has sought to determine the effects of accent on listening comprehension and external factors (such as familiarity with the target accent) that affect this dynamic. The definition and operationalization of *accent* can be a bit of a moving target, because it contains a subjective as well as objective component. Derwing and Munro (2005) acknowledged this relativity when they defined it as the extent to which an L2 learner's speech is perceived to differ from a target variety; Lippi-Green (2011 pp. 44–45) echoed this by saying that "accent can only be understood and defined if there is something to compare it with." Harding (2011) included more concrete phonological components when he indicated that accents comprised differences in the segmental and suprasegmental features of pronunciation, including variation in vowels and consonant sounds as well as stress and intonation. For the purposes of this study, we support Ockey and French's (2016, p. 695) construct definition of *accent*: "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."

Studies have shown that even a perceived accent on the part of the listener can affect comprehension of the speaker (Kang, 2012; Rubin, 1992). Ahn and Moore's (2011) investigation of listening comprehension of accented speech found that attitudes regarding specific nonnative English accents directly affected comprehension scores. In their study, nearly 200 university students were asked to complete an attitudes questionnaire that focused on accents before they performed an instruction-based assessment delivered in one of five accents (mild German, heavy German, mild Korean, heavy Korean, and native English speaker). Although the researchers found no significant differences in scores dependent on accent alone, they found that accent preferences served as a confounding variable on students' scores. More specifically, when groups were divided into low or high scores for accent perceptions, it was found that those who rated the Asian accents unfavorably scored lower on the comprehension assessment.

Attitudes can be at odds with quantitative data as well. Abeywickrama (2013) showed that the use of native and nonnative English accents as listening stimuli did not significantly affect nonnative listeners' comprehension scores on a retired TOEFL exam, indicating the use of accents did not impede their performances. Despite this, Abeywickrama still questioned the ethics of integrating nonnative English accents into high-stakes assessments. This was because, even though test takers had difficulty knowing which lectures were spoken by native and which by nonnative English speakers, their qualitative responses demonstrated preference for American and British English because they "deem them to be easier to understand and/or because they are considered 'the standard'" (p. 68). Thus, although results indicated that including nonnative accents did not actually hamper their performance, test takers believed that it would.

Bent and Bradlow (2003) coined the term *interlanguage speech intelligibility benefit* (ISIB), a theory that posits that NNES listeners may better understand English that is spoken by a person with the same L1. This is hypothetically due to a priming of specific acoustic-phonetic features, such as vowel or consonant deviations, pausing, or rhythm, that both speaker and listener may share (Pickering, 2006). Major, Fitzmaurice, Bunta, and Balasubramanian's (2002) study explored this theory when investigating the extent to which native-English-speaking and English as a second language (ESL) listeners performed better on a version of the TOEFL listening comprehension test when the speaker shared their L1. The results were inconclusive with regard to the ISIB. Specifically, Spanish listeners in the study seemed to be impacted by their L1 origin, scoring native Spanish speakers higher than those of other L1 backgrounds. In addition, Chinese and Japanese listeners scored well with Spanish L1 speakers, possibly due to prosodic similarity in rhythm among Chinese, Japanese, and Spanish (i.e., the lack of vowel reduction). However, the authors of this study noted limitations (e.g., not accounting for item difficulty, not measuring the strength of accent) that may have generated or enhanced the appearance of an ISIB.

In contrast, other studies have found little support for the existence of any shared L1 effect on listeners' judgments on speech constructs (e.g., accent, comprehensibility; Kang, Vo, & Moran, 2016; MacKay, Flege, & Imai, 2006; Munro et al., 2006), and it is possible that there is even an interlanguage intelligibility detriment (Julkowska & Cebrian, 2015). For example, in one study, listeners showed moderate to high agreement on speech judgments regardless of their L1 background (Munro et al., 2006). In another study, few differences were found in the ratings of accented speech between NES and NNES listeners (MacKay et al., 2006). A recent study (Kang et al., 2016) explored how 240 listeners from diverse language backgrounds weighed phonetic parameters (i.e., segmental features such as consonants and vowels, and suprasegmental features such as word stress and sentence stress) differently when rating nonnative speakers' speech for intelligibility, comprehensibility, and accentedness. The results suggest that, although listeners of English perceived accented speech in different ways for individual categories, depending on factors such as their L1 and their English instruction backgrounds, their global ratings scores were not significantly different. However, it must be noted that none of these studies included a comprehension component.

Other studies suggest that degree of accent can affect listening comprehension. Thirty years ago, Anderson-Hsieh and Koehler (1988) demonstrated that listeners' comprehension scores were lowest for the Chinese-accented English speaker with the most salient accent as determined by Test of Spoken English (TSE) scores and highest for the native English speaker. More recently, Ockey and French's studies (Ockey & French, 2016; Ockey et al., 2016) have corroborated that idea. For example, Ockey and French (2016) developed a Strength of Accent Scale based on salience, comprehensibility, and additional processing time. Listeners who heard speakers with scores of 2.0 or weaker on this scale (indicating little accent) scored as well on the TOEFL-like monologic comprehension test as when they heard a native English speaker. Furthermore, they generally had worse comprehension scores as the strength of accent increased.

Shared L1s and strength of accent are not the only factors that influence listeners' comprehension. A large body of research has investigated the link between familiarity of the target accent and comprehension (see Adank, Evans, Stuart-Smith, & Scott, 2009; Adank & Janse, 2010; Derwing & Munro, 1997; Gass & Varonis, 1984), generally finding that the more familiar a listener is with an accent, the more accurate their comprehension of accented speech. For example, Gass and Varonis (1984) demonstrated that listeners' judgments of comprehensibility of nonnative speech, although not directly linked to comprehension, were affected by their language experience. They found that listeners' familiarity with the topic, accent, speaker, and L2 speech were strongly correlated with their ability to transcribe sentences (often a measure of intelligibility) and summarize the stories they heard. Additionally, the English proficiency level of listeners is another factor that may influence the extent to which listeners comprehend accented speakers. Preliminary research has shown that perceptions of accent-related difficulty appear to be more salient among higher proficiency test takers (Harding, 2011). However, this is an area in need of future research.

Thus, despite pressure from the applied linguistics community for test developers to include various English accents, high-stakes assessment has been slow to do so for several reasons. First, the exact degree of influence of the ISIB, if there is one, remains unclear (e.g., Ahn & Moore, 2011; Bent & Bradlow, 2003; Julkowska & Cebrian, 2015; Kang et al., 2016; Major et al., 2002; Munro et al., 2006). Second, the degree of speaker accent may have an effect on listener comprehension (Ockey & French, 2016), and our field, as yet, does not have an accurate, objective way to measure this construct. Third, familiarity of accent adds an additional factor (Adank et al., 2009; Adank & Janse, 2010; Derwing & Munro, 1997; Gass & Varonis, 1984), and this may be compounded by proficiency of test takers (Harding, 2011). Also, test-taker perceptions play an important role as seen in Abeywickrama (2013) and in other accent perception studies (Rubin, 1992; Rubin & Smith, 1990). Lastly, there are various logistical issues to consider, such as which accents to include, how many, and in which parts of the test.

It is clear, then, that further research must be conducted before test developers and teachers feel that they can create a more ecologically valid test (using different English accents) while maintaining fair conditions for test takers. The current study adds to this body of literature by considering the impact of speaker comprehensibility. Acknowledging the possibility of L1-related variance in listeners' judgments of particular English varieties, we recruited listeners from L1 backgrounds that matched those of the speakers. The current study was guided by the following research questions: (1) To what extent do different English accents impact listeners' comprehension scores in the TOEFL iBT listening test? (2) To what extent does the listener sharing the same English accent as the speaker affect their comprehension scores?

METHODS

Participants

Speakers. Following Kachru's (1992) World Englishes model, three speakers from each of six distinct English varieties were recruited, for a total of 18 test speakers. Three U.S. and three British speakers typified inner circle varieties (where English is the dominant and first language of the majority); three Indian and three non-Anglophone South African speakers represented outer circle varieties (where English is an official language but not necessarily the L1); and three Chinese and three Mexican (L1 Spanish) speakers represented expanding circle varieties (where English is a language of international communication). We also ensured that speakers of each English variety shared the same geographic and linguistic origins so that their pronunciation would be comparable. For example, all speakers of American English were from California, and all Chinese English speakers reported speaking a standard Beijing dialect of Mandarin. However, it should be noted that geographic region is not the only predictor of accent. In England, for example, accent is related to socioeconomic class, and South African accents are linked to L1, class, and ethnolinguistic group (Van der Walt, 2000). Potential speakers were listened to by the research team (who are all trained in phonology) to ascertain similarities in speech varieties. All speakers had experience teaching university classes in English, and all held graduate degrees.

Consulting Harding's (2011) method, non–inner circle speakers (i.e., those from outer or expanding circles) were selected from a larger pool (about 41, approximately 10 speakers per country) to ensure they were proficient in their variety of English while also possessing a noticeable non–inner circle accent. Thus, before arriving at our final selection of speakers, possible candidates were asked to record themselves reading a sample TOEFL listening comprehension passage (3–4 minutes long). The research team then independently listened to each of the sample passages and determined whether speakers met specific qualities that indicated their professionalism as NNES educators (see Major et al., 2002, p. 50).

Subsequently, each selected speaker's degree of accent and comprehensibility were assessed by the research team (three members) as well as five other raters who had received extensive formal education in applied linguistics. Those who rated did so independently by listening to speakers' recordings of a sample TOEFL listening passage. The ratings were then compared, and the research team looked for consistent scores in the target range. The interrater reliability, as measured by intraclass correlation coefficients among these eight raters for selected speakers, was 0.968. Accent and comprehensibility were rated on two separate 1–5 scales (1 = *little accent/easy to understand* and 5 = *heavy accent/difficult to understand*). Of the speakers from each non–inner circle country, those representing low, mid, and high points on both scales were chosen. We chose a range of comprehensibility and accent to assess how the relative nature of these constructs could affect listening comprehension.

In addition, because the target test takers would likely not be trained in linguistics, phonology, or phonetics, the research team piloted the perceptions of this demographic. The results of these novice raters' evaluations were intended to ascertain that (a) the outer and expanding circle speakers had noticeable nonnative accents; (b) there were clear distinctions between low, mid, and high comprehensibility speakers; and (c) the ratings for each level of comprehensibility was approximately equal across L1s. Accordingly, in order to determine whether novice raters' judgments of accent and comprehensibility corresponded to those of the expert raters, 48 novice raters (13 males and 35 females) were recruited in an undergraduate linguistics class, other graduate classes on campus, or via e-mail. These raters included U.S. undergraduate students (10), international undergraduate students (9), U.S. graduate students (15), international graduate students (6), and instructors (both NESs [3] and NNESs [5]). For each speaker, listeners were asked to complete two 7-point Likert scale items to reflect accent and comprehensibility (1 = *no accent*/7 = *heavy accent* and 1 = *easy to understand*/7 = *difficult to understand*). Their ratings were averaged and were found to be similar to the initial expert rankings. Because the English proficiency of most speakers selected for this study was relatively high as university instructors, the lowest mean scores of comprehensibility and accentedness were 5.04 and 5.01 out of 7, respectively. More details about this additional test result can be found in Kang, Thomson, and Moran (2018).

Because the speakers of American and British English were all native speakers of English and all of a standard variety, they were all judged to be highly comprehensible to the research team. Consequently, unlike the non–inner circle speakers, we expected no test score differentiation for test passages spoken by these speakers.

Listeners. Sixty listeners took part in the TOEFL iBT listening and comprehensibility tests. These included 10 speakers from each of the six L1 speech varieties represented in the listening materials (i.e., 10 American, 10 British, 10 Indian, 10 non-Anglophone South African, 10 Chinese, and 10 Mexican); the American and British listeners were all native speakers, and the nonnative listeners were all highly proficient in English (i.e., they had TOEFL iBT scores of 100 or higher). Most of the listeners were recruited (and were residing) in their home countries, although some were currently residing in English-dominant countries for educational reasons.

Highly proficient listeners were targeted for this study. Although controlling for proficiency level may limit the generalizability of the findings, it was necessary to provide a clearer interpretation of the data because listeners' proficiency levels can serve as a confounding variable. Research has shown that perceptions of speakers' comprehensibility may differ based on the proficiency of the listener (Harding, 2011). Low-proficiency listeners may score accented speakers lower on comprehensibility than do their higher proficiency counterparts; due to their stronger overall language proficiency, highly proficient listeners are likely to be better equipped to negotiate speaker characteristics as well as text features than are lower proficiency listeners, who in turn may be more dependent on phonological features to aid comprehension. Understanding the effects of incorporating international accents, as well as the effects of shared L1s, with a high-proficiency listener baseline is a crucial first step in this line of research. It is also the best starting point because these high-stakes tests are primarily intended to identify highly proficient speakers for the purposes of advanced study in English. Subsequent research by our research team will address the effect of accent on comprehension for less proficient listeners.

Prior to performing the listening tasks, listeners were asked to complete a short survey to obtain demographic information (e.g., age, gender, ethnicity, country of origin), language background, educational experience, and scores on the TOEFL or equivalent English language proficiency test. We also ascertained, through a yes/no question, that all participants had normal hearing. At the end of the survey was a short diagnostic test, which consisted of a one-passage listening test with six questions derived from a currently available TOEFL iBT practice test recorded by a standard American English voice. Listeners were instructed that, in order to be selected for the full listening tests, they must perform successfully on the practice selection with one or no mistakes/incorrect answers. This information, in addition to the listeners' high TOEFL scores, served as assurance regarding the participants' current listening proficiency. Two potential listeners were not able to participate due to this screening.

Materials and Recordings

In contrast with the IELTS, which has a four-part listening section comprising two conversations and two monologues, each with 10 corresponding questions (IELTS, 2017), the listening section of the TOEFL iBT includes four to six academic lecture excerpts (each 3–5 minutes in length) and two to three conversations related to typical university life. Test takers answer six questions per lecture and five questions per conversation in a span of 60–90 minutes. Images on a computer screen accompany the audio recording, indicating both the context of the recording and the number of speakers (Manhattan Review, 2017). Because we were focusing on one L1 accent at a time, we only used academic lecture excerpts in this study.

The 18 TOEFL passages used were mock lectures that contained an introduction, body, and conclusion (ETS, 2016) and designed to measure test takers' "ability to understand spoken English" (Educational Testing Service, 2012). They were selected based on the results of analyses of item difficulty, after incorporating information about the degree of difficulty (both actual and perceived) and familiarity completed by 45 TOEFL preparatory students (Kang et al., 2018). Each passage's questions usually comprised traditional multiple-choice questions with four answer choices and a single correct answer. However, other types of questions included multiple-choice questions with more than one answer, questions that required the test taker to put in order events or steps in a process, and questions in which the test taker had to match objects or text to categories in a table (Educational Testing Service, 2012).

The selected passages were assigned across all six groups based on the English variety spoken. Then, we commenced recording the test lecture materials. Speakers were asked to record themselves reading the assigned listening stimulus passages (3–5 minutes) from the TOEFL iBT listening texts of academic lectures. The form of the TOEFL type materials controlled for style and content (e.g., passages of 500–800 words, containing mostly monologic speech, where the professor does all of the talking, with six questions per lecture). All the listening passages assigned to each speaker were similar in style but different in content; that is, speakers recorded listening passages that were unique to them. All selected messages fell in the range of 0.41–0.49 for type token ratio, 75–81% for most common 1,000 words, and 3.7%–4.4% for the use of an academic word list. A member of the research team was present for some of the recordings but served only to record; others were recorded remotely and sent digitally.

Speech files were controlled for speech rate to avoid a rate effect on comprehensibility (Munro & Derwing, 2001). If speakers' speech fell outside the range of 2.2–2.8 words per second (approximately 3.2–3.6 syllables per second), they were asked to rerecord and were given a sample with a more ideal speech rate after which to model themselves. Before recording, speakers were asked to practice reading the full passages and to discuss any lexical or pronunciation issues they might have with the research team. If any noticeable hesitation or lexical substitutions occurred that could not be corrected through sound editing, the speaker was asked to read the passage again.

The edited sound files were then embedded into several surveys using the assessment tool SurveyGizmo (www.surveygizmo.com). All listening passages were presented in a randomized order. When the listeners had completed the file, they could move to the questions page by clicking "next." A completion status bar showed listeners how much of the survey they had completed.

Data Collection Procedures

Listeners completed listening comprehension tests with questions based on 18 TOEFL iBT texts, answering 108 questions about those minilectures (18 texts × 6 questions each). The listeners were instructed to complete the comprehension test on one day in one sitting (2 hours). On two subsequent days, participants completed additional tasks; they were compensated the equivalent of US$120 for their time.

The listeners were told that they were allowed to take notes and to listen to each listening comprehension task one time only. They were also asked to report their testing behaviors after their completion, which ensured the research team that the listeners would not disregard the directions. As an added measure, each listening session was highly controlled and supervised; a computer lab was reserved at each data collection location, and students brought their own headsets. A known and trusted contact was tasked with personally observing each session. Only after the test environment was approved were the test links sent to the listeners. Listeners' responses were coded manually as correct or incorrect. Listeners received a 1 for each correct response and a 0 for each incorrect response.

Data Analysis

The primary analysis used in this project was a linear mixed-effects design. Linear mixed-effects models include both random effects of subject and item and fixed effects of independent variables (Faraway, 2005). In general, linear mixed-effects models are more flexible and robust than general linear models (GLMs); linear mixed-effects models can handle an unbalanced design or missing data, which are not tolerable in GLMs. Accordingly, linear mixed-effects models with a balanced design were estimated to investigate the research questions concerning the effect of speakers' L1 and any shared L1 effect, using SPSS version 22. We treated each listener and speaker as random effects, and the listeners' L1 and shared L1 as fixed effects. Then, the shared L1 was computed as an interaction between the speaker's and listener's L1s. In the case of the current study, the linear mixed-effect models offered the interaction effect and its interpretation in a more simplified manner, compared to mixed factorial designs (e.g., six speakers × six listeners).

Due to the nature of the recruitment of listeners and as part of the initial test, we started with our effort to determine if there was any significant difference among the listener groups across different accents or any difference between NES listeners (i.e., inner circle listeners) and NNES listeners (i.e., outer circle and expanding circle listeners) in their listening comprehension scores. We performed additional univariate post hoc analyses for the listener groups and the speakers, respectively, to assist the interpretation of the data and to provide the selection process of participant groups for the final analyses.

RESULTS

Difference Among the Listener Groups and the Speaker

The listener groups. The linear mixed designs were computed using each listener and speaker as random effects and the listeners' L1 and shared L1 as a fixed effect. The shared L1 status was also calculated as part of this model. Table 1 reports the main effects of the listener's accent. The results show that the inner circle listeners performed significantly lower than the rest of the listener groups ($p <$ .001). The American listeners' performance was significantly lower, with estimates of .50 lower than the averaged estimates of 5.46. The British listeners demonstrated the similar pattern, with estimates of .45 lower than the averaged estimates of 5.46.

Descriptive statistics as shown in Table 2 confirm that the American and British listeners performed significantly less well than the other

TABLE 1
Estimates of Fixed (Main) Effects for the Listener's First Language

						95% confidence interval	
Parameter (listener)	Estimates	Std. error	*df*	*t*	Sig.	Lower bound	Upper bound
Intercept	5.468	.894	.000	6.112	.000	3.759	7.177
American	−0.500	.109	1073	−4.567	.000	−0.714	−0.285
British	−0.455	.109	1073	−4.161	.000	−0.670	−0.240
Indian	0.088	.109	1073	0.812	.417	−0.126	0.303
South African	0.022	.109	1073	0.203	.839	−0.192	0.237
Mexican	−0.166	.109	1073	−1.522	.128	−0.381	0.048
Chinese	0.175	.084	1073	1.679	.104	−0.342	0.009

TABLE 2
Group Mean Scores on Listening Comprehension Test by Listener's First Language for All 18 Speakers

			95% confidence interval	
Listener group	Mean	*SD*	Lower bound	Upper bound
American	4.822	1.299	4.631	5.013
British	4.867	1.115	4.702	5.031
Indian	5.411	0.837	5.287	5.534
South African	5.344	0.905	5.211	5.478
Mexican	5.156	1.102	4.993	5.318
Chinese	5.322	0.907	5.188	5.456
Total	5.154	1.063	5.090	5.217

listeners. The Tukey post hoc analysis showed that there was no significant difference in listening comprehension scores between the two inner circle listener groups ($p > .99$). Also, no significant differences were found among the four outer and expanding circle listener groups in their performances ($p > .18$).

Even though this result was not something we expected when we started our project, it was not surprising, particularly given that the recruitment process was vastly different between the inner circle listeners and the outer and expanding circle listeners. That is, all the outer and expanding circle (or NNES) listeners were recruited strictly based on their TOEFL iBT/paper-based test (PBT) scores. The minimum requirement of the listeners' TOEFL scores was 100 iBT or 600 PBT or higher. This means that the NNES listeners who participated in this project were highly proficient in English and very skillfully trained in taking the high-stakes proficiency test to achieve such high scores. In fact, many of the listener participants had TOEFL iBT scores above 110. On the other hand, the inner circle NES listeners had no experience in taking any of the language proficiency tests, much less the TOEFL listening test. Even though we required a practice test before the actual administration, the inner circle listeners' familiarity with the TOEFL listening test was extremely limited. In addition, (lack of) accent familiarity related to listening materials, particularly in the learning of English, could be another possible factor for explaining the performance of native speakers. Overall, the participants between the inner circle and the outer and expanding circles appeared to be by and large different in nature. However, it is important to note that these results do not indicate that native English speakers necessarily score lower on the TOEFL than nonnative English speakers, but merely that those in our sample did due to our recruitment procedures.

Accordingly, the two inner circle groups were not included in the subsequent analysis. Henceforth the listener groups refer to the outer circle and the expanding circle participants only. Excluding the NES listeners is well justified because those who take the TOEFL tests and use the scores are ultimately NNESs, not NESs.

The speakers. The linear mixed designs were computed to examine how the 40 listeners responded to the varieties of 18 speakers' English through listening comprehension tests. The model used each listener and speaker as a random effect and the speaker's accent and shared L1 as a fixed effect. The results show that there was a significant main effect for the speaker's accent ($p < .001$). That is, the American, British, and Indian speakers were associated with significantly higher listening comprehension scores than the South African, Mexican, and Chinese speakers ($p < .001$). Their estimates were .44, .41, and .40 higher than the average score of 4.96, respectively (see Table 3).

TABLE 3
Estimates of Fixed (Main) Effects for the Speaker's First Language

						95% confidence interval	
Parameter (listener)	Estimates	Std. error	*df*	*t*	Sig.	Lower boun	Upper bound
Intercept	4.96	.077	.000	63.652	.000	4.808	5.114
American	.444	.110	1074	4.032	.000	0.228	0.660
British	.411	.110	1074	3.730	.000	0.194	0.627
Indian	.400	.110	1074	3.629	.000	0.183	0.616
South African	−.072	.110	1074	−0.655	.512	−0.288	0.144
Mexican	−.038	.110	1074	−0.353	.724	−0.255	0.177
Chinese	−.088	.110	1074	−0.659	.510	−0.287	0.142

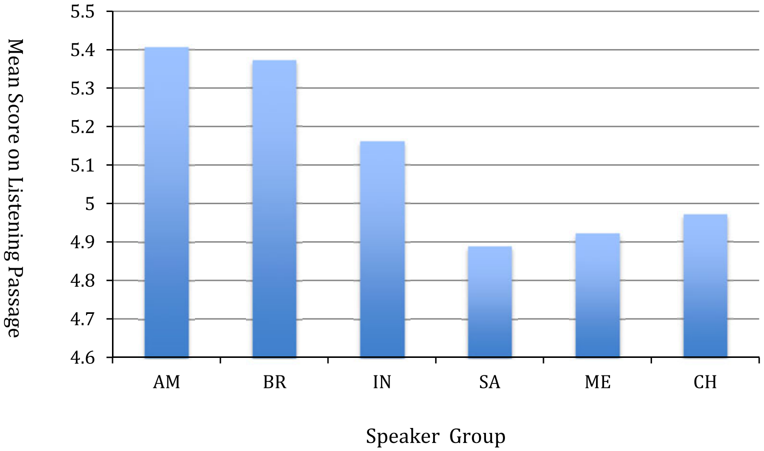

FIGURE 1. Speakers' mean scores on listening comprehension test for all 18 speakers. ***Note.*** **AM = American, BR = British, IN = Indian, SA = South African, ME = Mexican, and CH = Chinese. [Colour figure can be viewed at wileyonlinelibrary.com]**

TABLE 4

Group Mean Scores on Listening Comprehension Test by Speaker's First Language for All 18 Speakers

			95% confidence interval	
Listener group	Mean	*SD*	Lower bound	Upper bound
American	5.406	0.809	5.286	5.525
British	5.372	0.845	5.247	5.496
Indian	5.161	0.973	5.018	5.504
South African	4.888	1.290	4.697	5.079
Mexican	4.922	1.150	4.753	5.091
Chinese	4.972	1.085	4.812	5.131
Total	5.153	1.063	5.090	5.217

Including all 18 speakers that contained a wide range of comprehensibility, the American and British speakers received the highest scores followed by the Indian speakers (Figure 1 and Table 4). On the other hand, the post hoc analysis showed the listeners' performance was significantly lower with the other three speakers (South African, Mexican, and Chinese). There was no significant difference among the three low-scored accents ($p > .971$). The mean difference between the first three groups of speakers (AM, BR, and IN) and the remaining three groups of speakers (SA, ME, and CH) were significant ($p <$ .006), even though some lower bound scores may overlap with other upper bound scores at times (e.g., the lower bound of Indian [5.018] with the upper bound of South African [5.079]).

The Effect of Speakers' Accent and a Shared L1 Effect With All 18 Speakers

In order to answer the research question, we examined whether listeners performed better on a test of listening comprehension in English when the speaker shared the listener's L1 (see Table 5). Linear mixed-effects models were conducted to better understand the interaction between the language of the speaker and the language of the listener group. The first model examined the interaction of the speaker's accent and the shared L1 based on the listeners' L1 identity. This model was designed by including all 18 speakers, whose speech varied from low comprehensibility to high comprehensibility as determined by the trained raters and 48 additional listeners. Each speaker and listener was entered for the random effects. For the fixed effects, the model selected the speaker's accent and the shared L1 effect.

TABLE 5

Estimates of Fixed Effects for the Interaction of the Speaker's First Language and the Shared First Language Influence for All 18 Speakers

						95% confidence interval	
Parameter (interaction)	Estimates	Std. error	*df*	*t*	Sig.	Lower bound	Upper bound
Intercept	5.000	.168	710	29.810	.000	4.670	5.329
[SpeakerAccent = AM] × [accent-not-shared]	.517	.188	710	2.755	.006	0.148	0.884
[SpeakerAccent = BR] × [accent-not-shared]	.442	.188	710	2.355	.019	0.073	0.809
[SpeakerAccent = IN] × [accent-not-shared]	.489	.194	710	2.524	.012	0.108	0.869
[SpeakerAccent = IN] × [accent-shared]	.700	.237	710	2.951	.003	0.234	1.165
[SpeakerAccent = SA] × [accent-not-shared]	.100	.194	710	0.516	.606	−0.280	0.480
[SpeakerAccent = SA] × [accent-shared]	.633	.237	710	2.670	.008	0.167	1.099
[SpeakerAccent = ME] × [accent-not-shared]	.078	.194	710	0.402	.688	−0.302	0.458
[SpeakerAccent = ME] × [accent-shared]	−.367	.238	710	−1.546	.123	−0.832	0.099
[SpeakerAccent = CH] × [accent-not-shared]	.200	.194	710	1.033	.302	−0.180	0.580
[SpeakerAccent = CH] × [accent-shared]	.204	.211	710	1.247	.299	−0.160	0.590

Note. AM = American English; BR = British English; IN = Indian English; SA = South African English; ME = Mexican English; CH = Chinese English. The American and British listener groups were excluded from this model; accordingly, the interaction between the speaker's first language and the listeners' shared first language influence was not included.

When lectures were delivered by Indian speakers to listeners who did not share their accent, the listeners' performance was significantly different (p = .012). The listeners responded to the Indian accent more positively than some other accents, showing .489 higher estimates than the intercept. What is more, the Indian listeners performed significantly better on lectures delivered by Indian speakers (p = .003) than they did on lectures spoken by speakers with other accent backgrounds. The same pattern was found with the South African listeners. When the lectures were delivered by South African speakers, the South African listeners benefited significantly on their comprehension tests. The rest of the variables did not reveal any interaction effects.

The Effect of Speakers' L1 and Listeners' Shared L1 Status With Highly Comprehensible Speakers

The study aimed to explore the degree to which variables such as the speaker's L1 and the listener's shared L1 influence could interact to impact intelligibility. Some interaction effects were found with a varying degree of comprehensibility, which was initially determined by the trained raters. The study made a further step to investigate whether the interaction would exist if the listeners listened solely to the highly comprehensible speech, which was considered as both relative to the scale and relative to the other speakers of the same L1. Accordingly, the following linear mixed effect model included three speakers from each of the inner circle groups and one highly comprehensible speaker from each of the four countries representing the outer and expanding circles (India, South Africa, Mexico, and China). While examining the listening test scores, all three Indian (Hindi) speakers appeared to be somewhat more highly evaluated than the other three outer and expanding circle speaker groups. However, we chose only one speaker (rated most high) out of three in each country to make it consistent in the selection process and to adhere to the initial screening process. The inner circle speakers did not differ significantly in the listeners' performance; accordingly, all of them were included (see Table 6).

The linear mixed-effects model in Table 7 illustrates that none of the interactions showed a significant effect for the speaker's L1 and the shared L1 status for the highly comprehensible 10 speakers. In other words, as long as the speakers were highly comprehensible, the listeners did not show differences in their listening comprehension tests.

TABLE 6

Group Mean Scores on Listening Comprehension Test by Listeners' First Language for Highly Comprehensible 10 Speakers

			95% confidence interval	
Listener group	Mean	*SD*	Lower bound	Upper bound
American	5.516	.685	5.392	5.640
British	5.441	.797	5.297	5.585
Indian	5.475	.715	5.246	5.703
South African	5.500	.960	5.192	5.807
Mexican	5.550	.597	5.359	5.740
Chinese	5.600	.632	5.397	5.802
Total	5.500	.739	5.427	5.572

TABLE 7

Estimates of Fixed Effects for the Interaction of the Speaker's First Language and the Shared First Language Status for Highly Comprehensible 10 Speakers

						95% confidence interval	
Parameter (interaction)	Estimates	Std. error	*df*	*t*	Sig.	Lower bound	Upper bound
Intercept	5.700	.564	.000	10.100	0.000	2.7987	8.601
[SpeakerAccent = AM] × [accent-not-shared]	−0.183	.243	390	−0.752	0.452	−0.662	0.295
[SpeakerAccent = BR] × [accent-not-shared]	−0.258	.243	390	−1.060	0.290	−0.737	0.220
[SpeakerAccent = IN] × [accent-not-shared]	−0.300	.270	390	−1.110	0.268	−0.832	0.231
[SpeakerAccent = IN] × [accent-shared]	0.155	.331	390	0.000	1.000	−0.650	0.650
[SpeakerAccent = SA] × [accent-not-shared]	−0.333	.270	390	−1.233	0.218	−0.864	0.198
[SpeakerAccent = SA] × [accent-shared]	0.200	.331	390	0.604	0.546	−0.450	0.850
[SpeakerAccent = ME] × [accent-not-shared]	−0.100	.270	390	−0.370	0.712	−0.631	0.431
[SpeakerAccent = ME] × [accent-shared]	−0.300	.331	390	−0.906	0.365	−0.950	0.350
[SpeakerAccent = CH] × [accent-not-shared]	−0.133	.270	390	−0.493	0.622	−0.664	0.398
[SpeakerAccent = CH] × [accent-shared]	0.102	.281	390	0.301	0.678	−0.560	0.740

Note. AM = American English; BR = British English; IN = Indian English; SA = South African English; ME = Mexican English; CH = Chinese English. The American and British listener groups were excluded from this model; accordingly, the interaction between the speaker's first language and listeners' shared first language status was not included.

DISCUSSION

Overall, we found that listeners who were speakers of outer and expanding circle varieties attained significantly better scores on iBT listening passages produced with American, British, or Indian English accents, relative to those passages produced with South African, Mexican, or Chinese English accents. However, we also found that TOEFL listening passages produced with inner circle accent varieties resulted in higher scores than those produced with an Indian English accent. In sum, when degree of speaker comprehensibility was excluded as a factor, the following hierarchy emerged with respect to the comprehension of materials presented in the target English accents, from highest comprehension scores to lowest comprehension scores: American, British > Indian > Chinese, Mexican, South African. (Note that in this study, the > symbol indicates a significant difference.)

Listeners' stronger performances on materials spoken with American or British accents might be explained on the basis of listener familiarity with particular accents (see Gass & Varonis, 1984). For example, it can reasonably be assumed that the participants in our study would have had the greatest exposure to these two most dominant models, whereas their experience with the other varieties would be much less, except in cases where a given listener was a speaker of one of those varieties (see Major, Fitzmaurice, Bunta, & Balasubramanian, 2005). The only contradiction to this pattern was the listeners' relatively superior comprehension of passages produced with an Indian accent. The unique Indian English effect may stem from its greater phonetic similarity to British English relative to the other varieties represented. Indian English pronunciation evolved as a variety out of an attempt by its early users to acquire British RP. The influences of RP, with which listeners are likely more familiar, are still apparent in Indian English (see Pandey, 2015). Thus, listeners who are comfortable listening to RP might be able to bootstrap on this experience when processing the Indian English variety. The relationship between phonetic similarity and ease of processing has previously been described by Major et al. (2002). Ultimately, empirical research is needed to determine if this reasonable hypothesis is true. In fact, further research could rank other World English accents in terms of their similarity to the listeners' L1 as another variable in determining what samples to use in listening tests adapted for specific learners as a way of further leveling the playing field. More research is required regarding listeners' reactions to different varieties of accents in World Englishes.

We also examined the extent to which listeners who shared the same English accent as the speaker in the listening practice would have better comprehension relative to listeners from a different accent group. Rather than finding any definitive support for Bent and Bradlow's (2003) interlanguage speech intelligibility benefit, we found that listeners always performed better when they heard the listening passage spoken with an American accent, regardless of the identity of their own accent. Furthermore, there was no shared L1 benefit for Chinese and Mexican speakers. In support of Bent and Bradlow, however, although Indian and South African listeners, both from outer circle countries, did best in response to an American accent, their listening scores evidenced a secondary preference for passages spoken with their own accents.

Although there was some indication of an interaction between the listeners' L1 and the relative comprehensibility of the speech sample, we found that the general lack of an L1 speech intelligibility benefit was robust for the highly comprehensible speaker from each accent group, demonstrating a mismatched interlanguage speech intelligibility benefit (Bent & Bradlow, 2003). As long as the speech was highly comprehensible, listening comprehension scores did not differ regardless of their L1s. In fact, this result is promising because it suggests the potential of incorporating different varieties of highly comprehensible non-Anglophone speakers in listening comprehension tests, which can support the issue of ecological validity discussed earlier in the article.

Taken together, these results indicate a complex interplay between listeners' familiarity and experience with particular accents, the phonetic/phonological similarity between unfamiliar accents and familiar accents, and the listeners' own accent. There is also likely an interaction between listeners' scores and passage content, because some individual lexical items may be more severely impacted by a particular difference in accent than are others. For example, the word *mill*, which was frequent in a passage by one of the Chinese speakers, was pronounced [mɪʊ]. Another accent may have pronounced this word more similarly to a native-like pronunciation, resulting in greater comprehensibility. This latter point is far beyond the scope of the current study, but worthy of further investigation nevertheless.

The results of our study indicate that many speakers of inner circle varieties of English (e.g., RP, GA), which have long been considered standards to which English learners aspire, may not themselves be capable of obtaining TOEFL-type listening comprehension scores on par with high-proficiency English speakers from putatively nonstandard outer circle varieties of English (e.g., Indian English), nor with high proficiency L2 English speakers from expanding circle countries (e.g., L1 Spanish speakers from Mexico). Note that in our study, the outer and expanding circle speakers were selected using high proficiency as a criterion. Controlling for the proficiency of test takers as we did in our study seems to suggest a practice effect, which upon reflection we do not find particularly surprising.

Ultimately, although these results raise questions about the ecological validity of the test, previous research demonstrates that the TOEFL has strong criterion-related validity in that it accurately predicts academic success by nonnative speakers of English (Sawaki & Nissan, 2009), which is its primary purpose. Given the fact that speakers of inner circle varieties of English will not take the TOEFL exam, the rest of our examination of our first research question focused on comprehension of material produced in varying accents, by listeners from outer and expanding circle varieties only. These groups represent real-world examinees in TESOL fields. That is, we focused on outer and expanding circle listeners' performance on TOEFL listening passages spoken with accents representing inner, outer, and expanding circle varieties.

CONCLUSION

The current study aimed to provide guidance to promote the listening assessment as a test of international English and to help test practitioners understand the impact of different varieties of English on listeners' comprehension scores in high-stakes tests. Note that the current study corresponds to a long-aspiring desire expressed by many TOEFL test takers who may hold nonnative English accents themselves. When listeners took the comprehension test from the passages recorded by speakers with a various range of comprehensibility, there was a shared L1 effect. Indian and South African listeners benefited from their own accent when they listened to the listening comprehension passage. However, when listeners heard the highly comprehensible speech only, no shared L1 effect was found among the listener groups.

An attempt to use highly comprehensible World Englishes speakers can reduce the potential for unequal construct representation across groups and can have greater face validity with stakeholders (Harding, 2012). Importantly, listeners generally performed significantly better when they listened to prestigious English models such as Standard American English or British English (i.e., RP) as opposed to other models of World Englishes. Such findings support the results of earlier studies (Major et al., 2002, 2005) and offer empirical validation to TOEFL that including varieties of World Englishes on an English as a foreign language (EFL) listening test runs the risk of differentially advantaging and disadvantaging test takers from particular language backgrounds. It also further supports the argument that a listening test including World English varieties, in the interest of authenticity and internationalization, may create another form of test bias, thereby posing a threat to validity (Major et al., 2002). However, we discovered an important caveat: There was no significant difference among the listener groups when they listened to lectures delivered by the highly comprehensible speakers.

Findings also help the fields of language testing and L2 pronunciation understand the construct of listening comprehension in a more context-specific way: (a) general comprehension in speech communication and (b) comprehension in listening assessment. Finally, the findings offer an idea to answer an ongoing question of whether international tests of English proficiency should or should not privilege a standard variety of English to make it fair to speakers of nonstandard varieties (Hamp-Lyons & Davies, 2008).

In conclusion, the current study has limited generalizability because of the restricted number of World Englishes chosen from within each circle. It is true that speakers selected in this study may not represent the varieties of English at issue. Relatedly, different regional varieties of North American and British English can be included. Also, using six different L1s for listeners is not representative. Moreover, part of the results could be attributable to the familiarity of the lecture topics rather than to the speakers' accents even though we attempted to control for this issue when selecting the listening passages. Future research can be conducted to examine the interaction of test items, speakers' accents, and listeners, as well as the relationship between the familiarity effect and accent varieties. Finally, it must be noted that our results were based on scripted speech that was read aloud, which has different characteristics than conversational speech. Further research should investigate shared L1 effects on comprehension when listeners encounter authentic speech production; the present study has ecological validity only in the context of a TOEFL or similar listening test.

ACKNOWLEDGMENT

This research was funded by the Educational Testing Service (ETS) under a Committee of Examiners and the Test of English as a Foreign Language research grant. ETS does not discount or endorse the methodology, results, implications, or opinions presented by the researchers.

THE AUTHORS

Okim Kang is an associate professor in the Applied Linguistics Program at Northern Arizona University. Her research interests are speech production and perception, L2 pronunciation and intelligibility, L2 oral assessment and testing, automated scoring and speech recognition, World Englishes, and language attitude.

Ron Thomson is a professor of applied linguistics at Brock University. His research interests focus on the development of oral skills by L2 English learners. He is also the creator of www.englishaccentcoach.com, a free high variability phonetic training application for learning to perceive English vowels and consonants.

Meghan Moran is an instructor in the Applied Linguistics Program at Northern Arizona University. Her research interests include speech production and perception, L2 pronunciation and intelligibility, language planning and policy, language education policy, and linguistic discrimination.

REFERENCES

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

Adank, P., Evans, B., Stuart-Smith, J., & Scott, S. (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *Journal of Experimental Psychology*, *35*, 520–529. https://doi.org/10.1037/a0013552

Adank, P., & Janse, E. (2010). Comprehension of a novel accent by young and older listeners. *Psychology and Aging*, *25*, 736–740. https://doi.org/10.1037/a0020054

Ahn, J., & Moore, D. (2011). The relationship between students' accent perception and accented voice instructions and its effect on students' achievement in an interactive multimedia environment. *Journal of Educational Multimedia and Hypermedia*, *20*, 319–335.

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

Bent, T., & Bradlow, A. R. (2003). The interlanguage speech intelligibility benefit. *Journal of the Acoustical Society of America*, *114*, 1600–1610. https://doi.org/10.1121/1.1603234

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, *20*, 1–16. https://doi.org/10.1017/S0272263197001010

Derwing, T., & Munro, M. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly*, *39*, 379–397. https://doi.org/10.2307/3588486

Educational Testing Service. (2012). *TOEFL test prep planner*. Retrieved from https://www.ets.org/s/toefl/pdf/toefl_student_test_prep_planner.pdf

Elder, C., & Harding, L. (2008). Language testing and English as an international language. *Australian Review of Applied Linguistics*, *21*, 34.1–34.11. https://doi.org/10.2104/aral0834

ETS. (2016). The TOEFL iBT test: Improving your listening skills. Retrieved from https://www.ets.org/toefl/ibt/scores/improve/advice_listening_high

Faraway, J. J. (2005). *Extending the linear model with R*. Boca Raton, FL: Chapman and Hall/CRC.

Fitch, F., & Morgan, S. E. (2003). "Not a lick of English": Constructing the ITA identity through student narratives. *Communication Education*, *52*, 297–310. https://doi.org/10.1080/0363452032000156262

Gass, S. M., & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning*, *34*, 65–89. https://doi.org/10.1111/j.1467-1770.1984.tb00996.x

Hamp-Lyons, L., & Davies, A. (2008). The English of English tests: Bias revisited. *World Englishes*, *27*, 27–39. https://doi.org/10.1111/j.1467-971X.2008.00534.x

Harding, L. (2011). *Accent and listening assessment: A validation study of the use of speakers with L2 accents on an academic English listening test*. Frankfurt, Germany: Peter Lang.

Harding, L. (2012). Accent, listening assessment and the potential for a shared L1 advantage: A DIF perspective. *Language Testing*, *29*(2), 163–180. https://doi.org/10.1177/0265532211421161

Hyltenstam, K., & Abrahamsson, N. (2000). Who can become native-like in a second language? All, some, or none? *Studia Linguistica*, *54*(2), 150–166. https://doi.org/10.1111/1467-9582.00056

IELTS. (2017). Test format in detail. Retrieved from https://www.IELTS.org/en-us/about-the-test/test-format-in-detail

Jenkins, J. (2006). The spread of EIL: A testing time for testers. *English Language Teaching Journal*, *60*(1), 42–50. https://doi.org/10.1093/elt/cci080

Julkowska, I., & Cebrian, J. (2015). Effects of listener factors and stimulus properties on the intelligibility, comprehensibility and foreign accentedness of L2 speech. *Journal of Second Language Pronunciation*, *1*(2), 211–237. https://doi.org/10.1075/jslp.1.2.04jul

Kachru, B. B. (1992). *The other tongue: English across cultures*. Urbana: University of Illinois Press.

Kang, O. (2012). Impact of rater characteristics and prosodic features of speaker accentedness on ratings of international teaching assistants' oral performance. *Language Assessment Quarterly*, *9*, 249–269. https://doi.org/10.1080/15434303.2011.642631

Kang, O., Thomson, R. I., & Moran, M. (2018). Empirical approaches to measuring the intelligibility of different varieties of English in predicting listener comprehension. *Language Learning*, *68*, 115–146. https://doi.org/10.1111/lang.12270

Kang, O., Vo, S. T., & Moran, M. K. (2016). Perceptual judgments of accented speech by listeners from different first language backgrounds. *TESL-EJ*, *20*(1), 1–25.

Lippi-Green, R. (2011). *English with an accent: Language, ideology, and discrimination in the United States*. New York, NY: Routledge.

Llurda, E. (2004). Non-native-speaker teachers and English as an international language. *International Journal of Applied Linguistics*, *14*, 314–323. https://doi.org/10.1111/j.1473-4192.2004.00068.x

MacKay, I. R. A., Flege, J. E., & Imai, S. (2006). Evaluating the effects of chronological age and sentence duration on degree of perceived of foreign accent. *Applied Psycholinguistics*, *27*, 157–183. https://doi.org/doi.org10.1017/s0142716406060231

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*, 173–190. https://doi.org/10.2307/3588329

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

Manhattan Review. (2017). Listening Section of the TOEFL iBT. Retrieved from http://www.manhattanreview.com/toefl-listening/

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

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

Ockey, G. J., & French, R. (2016). From one to multiple accents on a test of L2 listening comprehension. *Applied Linguistics*, *37*, 693–715. https://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. https://doi.org/doi.org10.1080/10904018.2015.1056877

OMICS International. (2013). Ethnologue (17th ed.). Retrieved from http://research.omicsgroup.org/index.php/List_of_languages_by_total_number_of_speakers

Pandey, P. (2015). Indian English pronunciation. In M. Reed & J. Levis (Eds.), *The handbook of pronunciation* (pp. 301–319). Hoboken, NJ: Wiley.

Pickering, L. (2006). Current research on intelligibility in English as a lingua franca. *Annual Review of Applied Linguistics*, *26*, 219–233. https://doi.org/10.1017/S0267190506000110

Rubin, D. L. (1992). Nonlanguage factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants. *Research in Higher Education*, *33*, 511–531. https://doi.org/10.1007/BF00973770

Rubin, D. L., & Smith, K. A. (1990). Effects of accent, ethnicity, and lecture topic on undergraduates' perceptions of non-native English speaking teaching assistants. *International Journal of Intercultural Relations*, *14*, 337–353. https://doi.org/10.1016/0147-1767(90)90019-S

Sawaki, Y., & Nissan, S. (2009). *Criterion-related validity of the TOEFL iBT listening section* (ETS Research Report 2009[1]). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2009.tb02159.x

Taylor, L. (2006). The changing landscape of English: Implications for language assessment. *English Language Teaching Journal*, *60*(1), 51–60. https://doi.org/10.1093/elt/cci081

Taylor, L., & Geranpayeh, A. (2011). Assessing listening for academic purposes: Defining and operationalizing the test construct. *Journal of English for Academic Purposes*, *10*(2), 89–101. https://doi.org/10.1016/j.jeap.2011.03.002

Van der Walt, C. (2000). The international comprehensibility of varieties of South African English. *World Englishes*, *19*, 139–153. https://doi.org/10.1111/1467-971X.00165