Article

LANGUAGE TESTING

Language Testing
2024, Vol. 41(2) 263–289

DOI: 10.1177/02655322231179134
journals.sagepub.com/home/ltj


Fairness of using different English accents: The effect of shared L1s in listening tasks of the Duolingo English test

Okim Kang
Northern Arizona University, USA

Xun Yan
University of Illinois at Urbana–Champaign, USA
Beckman Institute for Advanced Science and Technology, USA

Maria Kostromitina
Northern Arizona University, USA

Ron Thomson
Brock University, Canada

Talia Isaacs
University College London, UK

Abstract

This study aimed to answer an ongoing validity question related to the use of nonstandard English accents in international tests of English proficiency and associated issues of test fairness. More specifically, we examined (1) the extent to which different or shared English accents had an impact on listeners' performances on the Duolingo listening tests and (2) the extent to which different English accents affected listeners' performances on two different task types. Speakers from four interlanguage English accent varieties (Chinese, Spanish, Indian English [Hindi], and Korean) produced speech samples for "yes/no" vocabulary and dictation Duolingo listening tasks. Listeners who spoke with these same four English accents were then recruited to take the Duolingo listening test items. Results suggested that there is a shared first language (L1) benefit effect overall, with comparable test scores between shared-L1 and inner-circle L1 accents, and

Corresponding author:
Okim Kang, Department of English, Arizona University, Liberal Arts Building 18, Room 140, Flagstaff, AZ 86011-6032, USA.
Email: okim.kang@nau.edu

no significant differences in listeners' listening performance scores across highly intelligible accent varieties. No task type effect was found. The findings provide guidance to better understand fairness, equality, and practicality of designing and administering high-stakes English tests targeting a diversity of accents.

Keywords
Accent varieties, assessing listening, attitudes, Global Englishes, listening tasks

Introduction

As of late, assessment researchers have been particularly interested in social and political aspects of language testing as it relates to fairness and justice and their relationship to test validity (Kunnan, 2014). International tests of English proficiency have been criticized on the grounds that such tests privilege a *standard* variety of English and are, therefore, unfair to speakers of nonstandard varieties (Hamp-Lyons & Davies, 2008). Scholars have supported the adoption of an English-as-an-International-Language approach over reference to traditionally standard varieties in international English proficiency tests (e.g., Taylor, 2006). Especially relevant to this movement toward Global Englishes (GEs) (Galloway & Rose, 2015; Rose et al., 2021), which encompasses the fields of World Englishes (WEs), English as a Lingua Franca (ELF), and English as an international language, is the assessment of listening skills. Because English instructors are from all around the world (Kang & Moran, 2019), an ecologically valid test of English listening would require listeners to be able to understand speakers with varied accents representing GEs.

The listening tasks of the Duolingo English Test (DET) include "yes/no" vocabulary and dictation tests. Although the association between these tasks and global listening skills has been well reported (e.g., Beeckmans et al., 2001; Nation & Newton, 2009), speaker characteristics in the listening stimuli can potentially incorporate greater variation. While previous research has examined the use of varied English accents in the TOEFL iBT (Kang et al., 2019; Ockey & French, 2014), in the IELTS tests (e.g., Kang et al., 2021), or in local English tests, such as at the University of Hawaii (Nishizawa, 2023), it is not yet known how different GE varieties can interplay with the listening tasks specific to the DET. The aim of the present project is, therefore, to explore the fairness of using different English accents in DET listening tasks by examining the impact of variability in accent on listener performance and any effect of the listener sharing the same English accent as the speaker in the prompt.

Note that in this paper, we use the terms of GE and WEs somewhat interchangeably, and also use native speaker (NS) and non-native speaker (NNS) in lieu of alternative wordings to avoid less transparent or long-winded alternative labels and to reflect the continued prevalence of those terms in some areas within applied linguistics (Isaacs & Rose, 2022; Kang et al., 2021); however, this in no way reflects an endorsement of native speakerism. In the data analysis and results sections, we categorize our speaker and listener groups as inner, outer, and expanding circle following Kachru's (1985, 1992) WE model to facilitate our data interpretation.

Literature review

The use of different English accents in listening assessment

Examining test-takers' reactions to globalization can help understand their lived experiences of validity which may lead to a more socially responsive enactment of language testing and assessment (Hamid et al., 2019). This effort is a part of evidence-based validity arguments, and it addresses assessment fairness by considering whether test-takers are tested in essentially the same way under the same conditions, or whether score interpretations and test-based decisions are equally appropriate for all test-takers (Kunnan, 2008). Given that English is the predominant second language (L2) spoken across the world (Eberhard et al., 2022), the role of test speakers' first language (L1) accent has been seen as a multifaceted area of inquiry in the assessment of L2 English listening skills (Llurda, 2004), especially under consideration of the WE framework (Kachru, 1985).

Perhaps the most well-known model describing GE use is Kachru's (1985) WE model that groups English varieties into three concentric circles: the inner circle (e.g., English spoken in the United States and the United Kingdom), the outer circle (English spoken in countries where it is an official language but not the language of day-to-day communication such as India and the Philippines), and expanding circle (English spoken in China or Mexico, where it is recognized as a Lingua Franca and is learned as a foreign language). However, since Kachru first put forth this model, much has changed both in the world and in the field of applied linguistics and language assessment. For example, as McArthur (2018) pointed out, in China (an expanding-circle country), English is used as a global language for various intra- and international purposes (see also Canagarajah, 2006), which calls for a more fine-grained discussion of local English varieties in this context (e.g., Davies, 2009; Jenkins et al., 2011).

From the testing perspective, the WE framework has brought another question related to inner-circle varieties (e.g., British and American English) and the reference to those as *standard* Englishes (Zhang, 2022). As Isaacs and Rose (2022) noted, the notion of a standard, globally accepted language is problematic especially since these English varieties are not used as widely as some L2 varieties. Although the label of "standard language" is unlikely to go away soon (Isaacs & Rose, 2022), this shift has motivated test developers to consider the validity issues of relying only on the inner-circle standard in language assessments (Brown, 2014; Harding & McNamara, 2018).

On one hand, there has been a push for the inclusion of L1 accent varieties in the listening sections of high-stakes tests. For example, Harding (2012) argued that listening tests need to reflect the reality of English learners encountering various L2 varieties if the ability to process L2 accents is included in the construct measured by a listening assessment. That is, if a test is not designed to measure this ability, and it is not part of real-world language use demands, the inclusion of L2 accents may be irrelevant. Further, discussing the issue of construct validity, Ockey and French (2016) suggested that the use of only one variety of English in a listening test is not representative of the measured construct, proposing that a test should include various accents but in a way that does not unfairly influence candidates' scores.

On the other hand, the complexities and potential drawbacks of featuring GE accents include the possibility of test bias, logistical concerns, and random error, resulting in test developers opting to preserve the status quo. Thus, Taylor and Geranpayeh (2011), while supporting the inclusion of accented varieties at advanced proficiency levels, opposed the introduction of what they called *accented* speech into testing at lower levels, due to the risk of depriving listeners of major phonetic cues necessary for listening comprehension. In addition, in parallel with earlier studies (e.g., Derwing & Munro, 1997; Gass & Varonis, 1984), they pointed out the issue of accent familiarity and exposure to certain accents in relation to bias in test-takers' results (see also Carey et al., 2011; Carey & Szocs, 2024). Besides the general issues of bias associated with accent exposure, several studies have shown that listening input presented in a nonfamiliar or new accent may disadvantage test-takers in comparison with a test where all items are recorded in a familiar L1 accent. Anderson-Hsieh and Koehler (1988) found that college-aged, American L1-English-speaking listeners' comprehension scores were significantly higher for passages that were recorded by speakers of American English rather than by Chinese English speakers. Similarly, Ockey and French (2016), in measuring listeners' comprehension of TOEFL iBT-based lectures recorded by L1 and L2 English speakers, showed that besides familiarity issues, even a light, less salient accent (as measured on the Strength of Accent scale developed by the authors) may affect comprehension scores. Moreover, test-takers from 148 countries in the study received lower scores on the items recorded by British and Australian speakers with relatively strong accents.

Related to the issue of accent familiarity is the notion of a shared-L1 intelligibility benefit. That is, when speakers and listeners share similar phonological systems in their L1, it leads to enhanced intelligibility (Bent & Bradlow, 2003). In describing this shared-L1 advantage, Bent and Bradlow (2003) coined the term "Matched Interlanguage Intelligibility Benefit" (p. 1606). The theoretical ground of this advantage is in the notion of L1 transfer, which affects a speaker's accent in an L2 and results in listeners from the same L1 possessing the same phonological patterns of that speaker's accent (Best, 1995).

Evidence of the positive role of shared-L1 and/or accent familiarity in judgments of accented speech has been ample, albeit inconsistent. Some studies have demonstrated that L2 listeners find speakers who share their L1 most comprehensible (Dai & Roever, 2019; Flowerdew, 1994; Saito et al., 2019). Furthermore, Harding (2012) applied differential item functioning (DIF) in investigating the potential for shared-L1 advantage in an academic English listening test and found that speakers from Chinese and Japanese language backgrounds demonstrated a shared-L1 advantage on some but not all items in the test, although the pattern was not clear for Japanese speakers. Shin et al. (2021) reinforced these findings in an analogous investigation of shared-L1 advantage for Chinese, Korean, and Indian test-takers using DIF analyses. Similar to Harding (2012), the observed shared-L1 effect was minimal and only for a few items and only for Chinese and Korean listeners. Other studies have further demonstrated that a shared-L1 advantage may hold only for certain languages (Kang et al., 2019; Major et al., 2002). For example, in Kang et al. (2019), listeners from India and South Africa performed better on a simulated TOEFL iBT listening test when they listened to a shared accent; however, listeners from expanding circles (Chinese and Mexican English) did not benefit from audio materials presented in a shared-L1 accent. Conversely, there are studies that have

found little evidence of a shared-L1 advantage for listening comprehension (Abeywickrama, 2013) and intelligibility (Munro et al., 2006), potentially because observed effects were possibly due to listeners' previous exposure to the accent rather than just shared L1 (Gass & Varonis, 1984; Munro et al., 2006; Ortmeyer & Boyle, 1985; Smith & Bisazza, 1982; Yule et al., 1990).

Despite the conflicting findings, a source of consensus is that the role of shared L1 decreases when listeners hear highly intelligible speakers (e.g., Kang et al., 2019; Ortmeyer & Boyle, 1985). In Bent and Bradlow (2003) and its replication by Stibbard and Lee (2006), highly proficient L2 English speakers were found to be as intelligible as L1 speakers by listeners from the same language background. In Kang et al. (2020), beginner, intermediate, and advanced listeners from South Korea completed listening comprehension and intelligibility tasks with GE accents. The results mirrored those in Kang et al. (2019), showing that as long as speakers were highly comprehensible and intelligible, advanced listeners performed equally well on the tests no matter the accent. However, results were more complex for intermediate listeners who showed significant differences in test scores across highly comprehensible speakers of different English varieties. That is, there was a small effect ($\eta^2 = .059$) of speakers' L1 on the intermediate listeners' performance on tasks, with, for this sample, a South African accent being significantly less intelligible than, for instance, a Mexican accent. Notably, intermediate and advanced listeners were also sensitive to less comprehensible GE speakers in the listening comprehension tasks, but not in the intelligibility task. Finally, studies (e.g., Kraljic et al., 2008) found that listeners' comprehension may be influenced by the idiosyncratic articulatory properties of speakers' and listeners' beliefs about speakers' identity including national origin (e.g., Niedzielski, 1999).

While several emerging trends can be observed based on the results of the studies discussed above, the effect of variably accented speech in listening tests is still largely unclear. For one, there is a lack of agreement on the shared-L1 advantage in listening comprehension. Findings regarding the overall inclusion of GE accents in listening tests are also inconsistent. Research is needed to validate the use of different English accents in listening assessment, particularly in the context of GE.

The interplay of different accents and listening tasks

As Buck (1994) noted, no listening test is "pure," as it often assesses additional language skills because test-takers are likely to complete reading or writing tasks integrated with listening. To date, a large variety of tasks have been used to assess listening skills, with some researchers and practitioners employing tasks based on communicative approaches, including dialogues and naturalistic conversations (Ross & Langille, 1997). Others have chosen to use integrative listening comprehension tests that focus on language processing for information, such as listening cloze, listening recall, and summary gap filling (Brown & Trace, 2018; Buck, 2001; Cai, 2013). Another such integrative listening task that has been widely used for assessment purposes is dictation. Earlier research showed that this task can act as a good supplement to other listening tests, as it builds on higher-order processes and can even serve as a communicative language test granted that the speech rate in the task is fast enough (Cohen, 1994; Weir, 1993). However, more recent

studies seem to suggest that dictation tasks use predominantly lower-level processes, as they do not require listeners to construct meaning or discourse (e.g., Jia & Hew, 2019). Regardless of their processing nature, dictation tasks have been found to be important predictors of learners' listening comprehension (Siegel & Siegel, 2015). In speech perception research, one can argue that this task is typically viewed as a measure of the speaker's intelligibility, since this construct is generally defined as a listener's ability to correctly transcribe words and sentences they hear (Kang et al., 2018, 2020); dictation is parallel to aural yes/no vocabulary tasks, as it is essentially a speaker's intelligibility measure, being that it requires speech processing at the phonemic level, knowledge of sound co-occurrences, and working memory involvement. Akin to dictation, this type of task has also been found a strong predictor of test-takers' listening comprehension (Harsch & Hartig, 2016; Matthews & Cheng, 2015; Milton et al., 2010). In addition, yes/no tests have been largely validated as a measure of vocabulary knowledge and for placement purposes (Harrington & Carey, 2009).

While these tasks are aimed at assessing listening comprehension objectively, task characteristics have been found to affect test-takers' scores (Wagner, 2014). One such characteristic is the input that listeners receive (Bachman & Palmer, 1996). Although the issue of WE accents in listening input does not appear to be extensively explored, recent studies have investigated the role of shared L1 in a function of varied listening tasks. Dai and Roever (2019) examined the effect of shared L1 in three sections of a listening test: sentence-based picture recognition, monologue-based true/false response, and monologue-based word gap filling. Learners in their study demonstrated varied performance on each task, with shared-L1 advantage being stronger in tasks where listeners needed to perceive specific words rather than understand propositions. With regard to intelligibility-based tasks in particular, robust empirical research has provided support for the idea that the use of outer- and expanding-circle English varieties does not negatively affect listeners' performance (Kang et al., 2020; Lagrou et al., 2013; Weber et al., 2011), especially with high-level listeners (Kang et al., 2019). As Field (2013) pointed out, small misarticulations by the speaker or mishearings by the listener at a phonemic level can be canceled out, since they are not likely to affect intelligibility. Kang et al. (2020) further clarified that if a speaker is highly intelligible (i.e., rarely exhibits segmental divergences that are consequential for understanding), their speech can be used for tasks like dictation or aural yes/no tasks, despite the presence of a mildly unfamiliar accent.

Because both the yes/no vocabulary and dictation tasks in the DET rely on phonemic speech processing, it can be argued that one essential characteristic that affects their usefulness for listening assessment is speakers' intelligibility, especially if outer- and expanding-circle English varieties were to be included in recorded prompts for the tasks. However, it is generally unknown how different GE varieties can interplay with these listening tasks. It is, thus, timely and important to investigate the validity of the use of GE accent varieties in the listening tasks of the DET and to explore the effect of accents on these task types.

This study

The aim of the present project was to explore the fairness of using different English accents in DET listening tasks by examining the impact of accent varieties on listeners'

performances and the effect of the listener sharing the same English accent as the speaker. We further investigated the effects of task types (i.e., "yes/no" vocabulary and dictation) on listeners' DET listening performance. The project was guided by the following research questions:

1. To what extent does listening to test materials spoken with a shared versus different English accent affect test-takers' performance on DET tasks targeting listening? Note that "shared" here refers to the listener sharing the same English accent as the speaker.
2. To what extent do different English accents affect listeners' performance on two DET listening task types (i.e., "yes/no" vocabulary and dictation)?

Methods

Participants

Speakers. A total of 24 speakers from six GE varieties provided recordings for the listening tasks. Two of the varieties were North American and British English (AmE and BE), and the other four included Chinese, Korean, Indian, and Mexican English, with Mandarin Chinese, Korean, Hindi, and Mexican Spanish being among the top 10 most frequent L1s of DET test-takers (LaFlair & Settles, 2020). Speaker selection was conducted by drawing on Harding's (2012) and Kang et al., (2019) methods. To ensure homogeneity of speakers, speakers of one GE variety had to be born and raised in the same geographic region. For North American English, speakers from the West Coast were selected; for British English, speakers from the South of England were selected; for Indian English, L1 speakers of Hindi; for Chinese English, L1 Mandarin speakers from Northeast China; for Mexican English, speakers of Spanish from Mexico; and finally for Korean English, speakers from South Korea. Initially, 77 speaker candidates for all GE varieties were recruited. Then, the six expert raters (i.e., applied linguists with PhD degrees) determined the speakers' intelligibility by transcribing five sample words, five nonword items, and five sentences from each. They also perceptually rated speakers' accentedness and comprehensibility on two distinct 5-point scales (1 = *very accented/extremely difficult to understand* and 5 = *not accented at all/completely easy to understand*). Expert raters' accentedness and comprehensibility ratings were averaged for each speaker. An average percentage of words transcribed accurately by the raters was calculated for each speaker's intelligibility. For the final speaker selection, a speaker's sample had to score at least four out of five or higher on accentedness and comprehensibility and at least 90% on intelligibility. The final selected speakers were asked to record the DET listening items.

Listeners. The listeners in the study were 160 learners of English from four L1 backgrounds that matched the sources of the DET's non-North American and non-British GE varieties; that is, the 160 learners grew up speaking Mandarin Chinese, Korean, Hindi, or Mexican Spanish (40 learners per group). They all had taken or had been planning to take an international English proficiency test within a year.

A total of 97 listeners were females, and 63 were male. The listeners' average age was 23.8 years (SD = 5.3 years), and their average length of English learning was 6.5 years. In

Table 1. Language varieties presented in the study.

Language variety	Standard vs. nonstandard languages	World Englishes
American English	Standard	Inner circle
British English	Standard	Inner circle
Indian English	Nonstandard	Outer circle
Chinese English	Nonstandard	Expanding circle
Korean English	Nonstandard	Expanding circle
Mexican English	Nonstandard	Expanding circle

terms of education, 146 participants were enrolled in bachelor's degrees, 12 were in master's programs, one was in high school, and one was in a PhD program. A total of 32 participants had lived in an English-speaking country outside of their own home country. Participants were also asked to share their proficiency test scores if they had taken such a test. For those who reported, scores ranged from 6.5 to 8.0 for IELTS, 95–135 for DET, 92–117 for TOEFL iBT, 470–640 for TOEFL ITP, and 845–990 for TOEIC.

Because this study makes use of several terms that exist in the field of assessment in reference to L2 English varieties as no current label reflects the construct perfectly, per Isaacs and Rose (2022), Table 1 provides mappings for each of the English varieties represented in the study's speaker and listener sample.

Recordings and materials

Each speaker produced recordings for 72 dictation items and 216-word items of equal difficulty that were provided by Duolingo. Item difficulty was checked and ensured by Duolingo in a process separate from this study. All recordings were made with high-fidelity audio equipment in a quiet room. The speakers were sent the items to record for the study together with the recording instructions and a sample recording of all the items (with a more ideal speech rate) made by a trained speaker. Before recording, speakers were encouraged to practice reading the items out loud and email about any lexical or pronunciation issues. To ensure the uniformity of the recordings across all speakers, the quality of all recordings was checked multiple times for absence of background noise and correct pronunciation of the items (e.g., word stress and voiced/voiceless segmentals). If any items had recording issues, they were re-recorded by the speakers. The recorded utterances were edited visually and aurally from a waveform display using PRAAT speech editing software so that the amplitude was normalized across the recordings. The speech rate of the speakers was also controlled to avoid a rate effect on comprehensibility (Kang et al., 2019). Thus, the recordings were adjusted as needed to make sure that they were consistent across the speakers ($M = 3.4$ syllables per second, $SD = 0.12$).

Instruments

Listener Background Questionnaire. Prior to performing the listening tasks, listeners completed a short survey to obtain demographic information, such as age, gender, ethnicity,

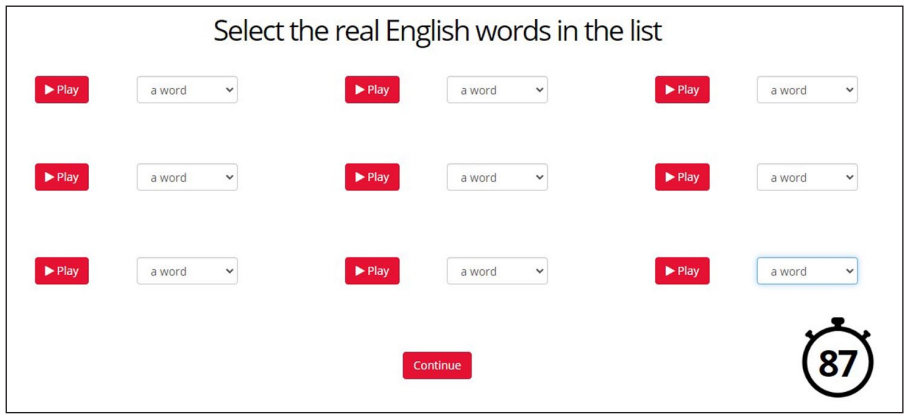

Figure 1. An example of a yes/no task.

country of origin, language background, and educational experience. The questionnaire included questions about listeners' experience with English learning and with English proficiency tests.

Listening tasks. The listeners completed the listening tasks in three phases. In the first phase, they took a yes/no vocabulary task and a dictation task in their choice of either American or British English. Listeners were given the option to choose between American and British varieties because those accents dominate in English leaners' learning materials. Listeners' preferences varied according to their geographic locations, which most likely corresponded with the accents most represented in their English-learning materials (Kang, 2015). In the second phase, listeners were offered the same two tasks but with their own L1 accent. Finally, in the third phase, listeners were led to a randomly assigned nontarget/not-shared accent. Each test included unique items; that is, dictation and yes/no items in one test were not repeated in the other two versions taken by a listener, and the order was randomized. Figures 1 and 2 below present screenshots demonstrating what the task looked like for the participants.

The design and presentation of the listening tasks mirrored that of the DET. That is, in the word recognition tasks, participants were presented with nine word/nonword items on one screen. They were allowed to listen to each item for an unlimited number of times, and they had 90 seconds to complete the task. The dictation items were individually presented, and participants were able to listen to each item for up to three times. Participants were given 60 seconds to type the transcription of each item.

Yes/no vocabulary task. Each of the three-word recognition tasks consisted of 18 existing/nonexisting English words (2 screens $\times$ 9 items on each screen, approximately 50% words were nonexisting in each task). For each participant and in each GE variety, the stimuli were randomly selected from the pool of items recorded by the four speakers from one variety. Although each task screen presented an uneven number of items ($N=9$), all four speakers' recordings were included in a balanced way with at least two recordings from each speaker per task.

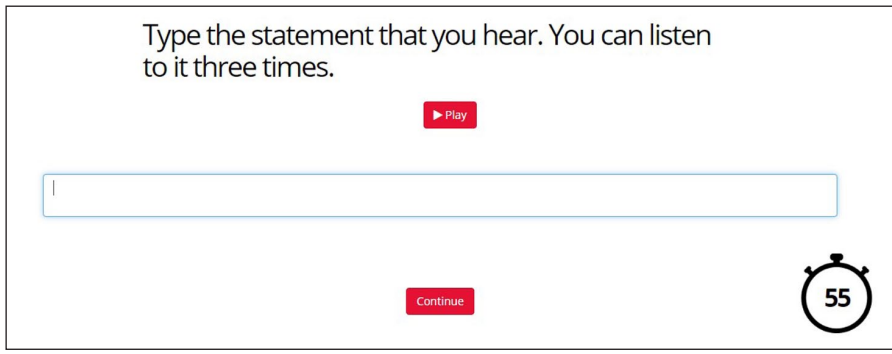

Figure 2. An example of a dictation task.

Sentence dictation task. Each of the three dictation tasks included eight items (four items × two times in a task). Similar to the word recognition task, the items in each task were chosen randomly from the pool of recordings. Voices of the four speakers in each GE variety were represented in a balanced manner (two sentences per speaker). However, to prevent the listener from hearing the same sentence in multiple tasks, the items in each sentence dictation task remained the same for all listeners who took the task.

Elicited imitation test. All participants took a 24-sentence elicited imitation test (EIT) as an independent measure of their general language proficiency within an academic context (Yan, 2020). EITs have been found valid in several studies that have examined the criterion and concurrent validity of the instrument (Kim et al., 2016; Kostromitina & Plonsky, 2022; Yan et al., 2016). In this task, participants were asked to listen to one sentence at a time, identify one of the two words that the sentence contained, and then repeat the sentence as accurately as possible. The stimuli in the task were assembled in three blocks based on their length as well as syntactic, phonological, and lexical complexity. While the blocks were presented in the same order to all participants, the stimuli within each block were presented randomly for each participant.

Procedure

The participants were given an option of choosing their target and L1 accents (Phases 1 and 2). We offered a choice regarding the target accent, since some test-takers are, we assume, predominately exposed to either British or American accents when they practice listening in classroom settings. We wanted the results of this study to be generalizable to existing widespread listening tests (e.g., TOEFL iBT and IELTS) that include both accents. After that, they were led to one randomly assigned (novel and not-shared) accent (Phase 3). These three phrases of accent procedures (i.e., the target accent → shared L1 accent → nonshared accent) were identical for all participants, that is, not randomized, for the purpose of exploring the participant's reactions to accent varieties in a structured manner. Data collection took place with the help of Gorilla Experiment Builder (Anwyl-Irvine et al., 2020). The participants were assigned unique IDs that they used to log in to the

platform. After logging in, they completed a background questionnaire which identified their L1 background. Depending on the test-taker's background, the experiment adjusted the task sequence to ensure that participants completed tasks with the shared GE variety in Phase 2 and that Phase 3 did not include their own GE variety. After the background questionnaire, participants completed the listening tasks followed by an EI task. In sum, 23 listeners chose to complete the task in Phase 1 with the British accents, and the remaining 137 listeners chose the American accent. For our primary analysis, these two sets of responses were combined as one L1 accent group; accordingly, this unbalanced sampling was not a concern for the study. In Phase 3, a version of the task with a nonshared GE accent was randomly assigned to each participant by the program presenting the tasks. A listener had an equal chance of completing the task with any of the three L2 accents with the exception of the shared-L1 accent. Thus, 38 participants listened to the Mandarin Chinese accent, 38 listened to the Korean accent, 37 to the Indian accent, and 47 to the Mexican Spanish accent in Phase 3. The entire sequence of tasks took on average 45 minutes to complete (range: 40–55 minutes).

Scoring

Followiong data collection, participants' responses for the listening task were compiled using a Python script, added to a CSV file and scored. A percentage of the correctly identified words and nonwords (i.e., true positives and true negatives) was calculated for the three yes/no vocabulary tasks completed by each participant. Given that there is little agreement about how the yes/no test should be scored (Pellicer-Sánchez & Schmitt, 2012), this simple approach was adopted due to our main interest in comparing test-takers' performance on the three versions of the tests with different accents rather than scoring their proficiency. Recent studies have also undertook the same scoring approach to a yes/no vocabulary test, (e.g., Kremmel & Schmitt, 2016; Zhang et al., 2019) further justifying our choice. To score the dictation tasks, a Python script was written to extract the transcriptions typed by the participants and compare them to the target transcription. The script employed *pandas* and *numpy* packages and calculated the percentage of words that were transcribed correctly in each sentence. Ten percent of the dictation items were scored by two raters to check the accuracy of the script and reached 98% agreement. A human rater completed the checking of the remaining transcriptions.

Statistical analysis

To address the first research question about the effect of shared versus different English accents in the DET listening, a series of multilevel linear mixed-effects models were conducted to examine whether participant and speaker backgrounds interacted to influence listeners' performance on the tasks. The models also included random factors of listeners and speakers. All analyses were run in R (ver.4.0.3). Before running the analyses, a residual plot was examined for evidence of assumptions violations. No such evidence was found. In addition, no multicollinearity was found in the independent variables (highest correlation $r = .25$). The following packages were used in the analyses: *backports* (Lang & R Core Team, 2020), *effects* (Bates et al., 2015; Fox, 2003; Fox & Weisberg, 2019; Long, 2019; Wickham, 2016), *lmerTest* (Kuznetsova et al., 2017), *psych*

Table 2. Summary of ANOVA results of listeners' performance on the listening tasks and EIT.

	df	*F*	*p*	Chinese (*n* = 40)		Korean (*n* = 40)		Indian (*n* = 40)		Mexican (*n* = 40)	
				M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
EIT	(3, 156)	4.467	.004	84.90	18.47	86.85	21.34	96.98	9.54	92.05	13.07
Yes/no task 1	(3, 156)	4.43	.005	0.75	0.13	0.81	0.14	0.71	0.15	0.80	0.13
Yes/no task 2	(3, 156)	4.341	.005	0.72	0.14	0.80	0.12	0.72	0.16	0.79	0.11
Yes/no task 3	(3, 156)	4.194	.006	0.69	0.14	0.78	0.14	0.69	0.17	0.77	0.11
Dictation task 1	(3, 156)	4.981	.003	0.74	0.14	0.86	0.14	0.82	0.12	0.80	0.13
Dictation task 2	(3, 156)	2.795	.042	0.79	0.13	0.85	0.14	0.79	0.10	0.77	0.13
Dictation task 3	(3, 156)	4.564	.004	0.70	0.15	0.80	0.16	0.78	0.14	0.80	0.12

ANOVA: analysis of variance; EIT: elicited imitation test; *SD*: standard deviation.

(Revelle, 2020), *plyr* (Wickham, 2011), *readxl* (Wickham & Bryan, 2019), *sjPlot* (Lüdecke, 2021), and *tidyverse*. To address the second research question about the task effects in the DET listening, another multilevel linear mixed-effects model was fit. The fixed factor of task was added to the model built to address the first research question.

Results

The results are presented in response to each of the research questions. To enhance the interpretability of the data, the speaker and listener groups were categorized into three: inner, outer, and expanding circle. First, the normality of the data was checked and confirmed with skewness and kurtosis all within the range of [−2, 2]. Table 2 includes descriptive statistics for scores on the EIT (as an independent measure of listener's language proficiency) as well as those on each vocabulary and dictation task on the DET across the four listener backgrounds. The EIT task demonstrated quite high internal consistency, as indicated by the associated Cronbach's alpha value ($\alpha = 0.84$, $SD = 0.05$, $SEM = 0.016$). Since vocabulary and dictation tasks included different items provided by Duolingo in each iteration of the task, it was not possible to calculate Cronbach's alpha for them; however, previous reports on the internal consistency of these items (e.g., LaFlair & Settles, 2020) indicated high internal consistency for the two tasks. To provide a general understanding of the data, we also included analysis of variance (ANOVA) results with participants' L1s as the grouping variable in Table 2 as a preliminary analysis without factoring out participants' proficiency. In terms of EIT scores (out of 120 points total), we observed a significant difference across listener background, with Indian listeners performing the highest, followed by Mexican listeners, and Chinese and Korean listeners performing the lowest. However, on the DET tasks, in general, Korean listeners and Mexican listeners outperformed Chinese and Indian listeners, with Korean listeners performing the highest. Note that there were three sets of yes/no tasks and three sets of dictation tasks included in the study by using different accents, with set 1 = inner-circle accent, 2 = shared-L1 accent, and 3 = other L2 varieties.

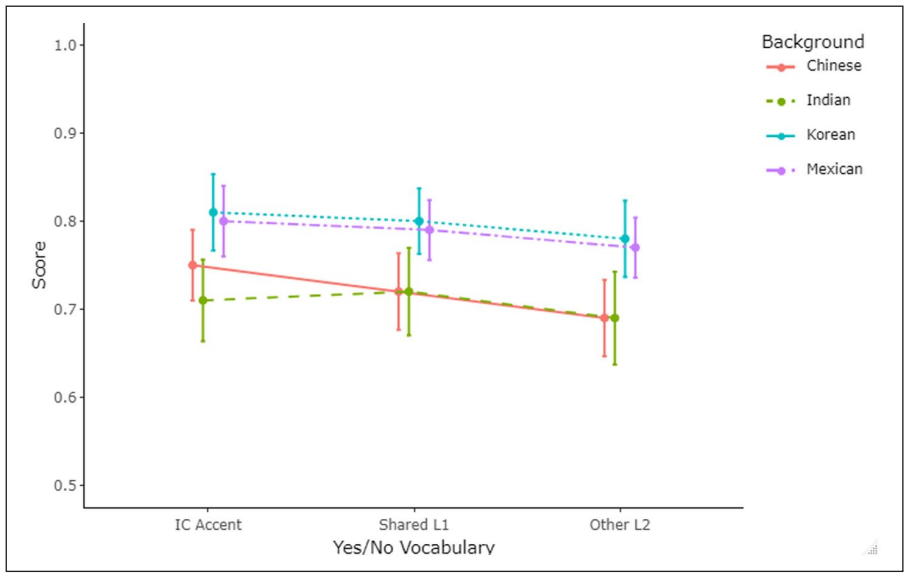

Figure 3. Listener performance on yes/no vocabulary tasks (means and 95% CIs).

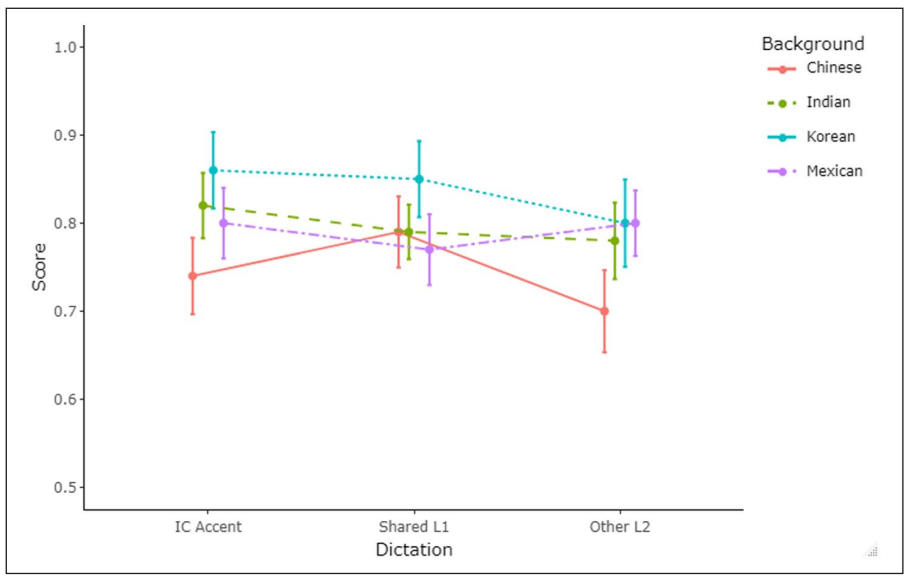

Figure 4. Listener performance on dictation tasks (means and 95% CIs).

Figure 3 shows the distribution of mean scores and associated 95% confidence intervals (CIs) for the task of yes/no vocabulary. Figure 4 details listeners' performance on the dictation tasks.

Table 3. Summary of the linear mixed-effects model with the best fit.

Predictors	Model 4		
	Estimates	CI	*p*
(intercept)	−0.028	[−0.065, 0.009]	.136
EIT	0.004	[0.003, 0.004]	<.001
Listener background: Korean	0.079	[0.043, 0.115]	<.001
Listener background: Indian	−0.021	[−0.060, 0.017]	.274
Listener background: Mexican	0.034	[−0.003, 0.070]	.071
Inner-circle accent	0.019	[−0.026, 0.063]	.416
Outer-circle (Indian) accent	−0.041	[−0.079, −0.003]	.035
Shared L1 accent	0.016	[0.000, 0.032]	.044
Random effects			
σ^2	0.03		
τ_{00}	0.005	Listener	
	0.002	Speaker	
ICC	0.214		
N	160	Listener	
	24	Speaker	
Observations	3564		

Note: Reference group for listener background = Chinese; reference group for speaker accent = other expanding-circle varieties. CI: confidence interval; EIT: elicited imitation test.

The effects of a shared versus different English accent on DET listening

Six models were tested, starting from a null model with only random intercepts for individual listener and individual speaker. Gradually, we added independent variables in the following sequence: main effect for standardized EIT scores (independent measure of listeners' general language proficiency), main effect for listener background (Chinese [reference], Korean, Indian, and Mexican), speaker background in concentric circles (inner-circle accent, outer-circle (Indian) accent, shared L1 accent, and other expanding-circle accents [reference]), and interaction between listener and speaker background. The response variable in the mixed-effects model was participants' scores on dictation and yes/no tasks. The best model in terms of overall fit was the model that included all random and fixed effects except the interaction between listener and speaker background. This model used Chinese as a reference group for listener background and selected expanding-circle varieties (Mexican, Chinese, and Korean) as a reference group for speaker accent. The fixed effects in the model explained over 11% of variance in listeners' scores (marginal $R^2 = .116$), and the random factors of listeners and speakers explained an additional 19% (conditional $R^2 = .305$). The model overview is given in Table 3, and the complete table with all models and their fit is included in Appendix 1. Due to no significance between speaker and listener background, interaction effects are not included in the table.

Based on the results of linear mixed-effects modeling, several observations were made about the data. First, Korean listeners outperformed the other listener groups on the

Table 4. Summary of the linear mixed-effects model with inner-circle varieties as the reference group.

Predictors	Model 4		
	Estimates	CI	*p*
(intercept)	−0.010	[−0.054, 0.035]	.675
EIT	0.004	[0.003, 0.004]	<.001
Listener background: Korean	0.079	[0.043, 0.115]	.001
Listener background: Indian	−0.021	[−0.060, 0.017]	.224
Listener background: Mexican	0.034	[−0.003, 0.070]	.455
Expanding-circle accent	−0.019	[−0.063, 0.026]	.416
Outer-circle (Indian) accent	−0.060	[−0.113, −0.007]	.033
Shared L1 accent	0.003	[−0.046, 0.041]	.910
Random effects			
σ^2	0.03		
τ_{00}	0.005	Listener	
	0.002	Speaker	
ICC	0.214		
N	160	Listener	
	24	Speaker	
Observations	3564		

Note: Reference group for listener background = Chinese; reference group for speaker accent = inner-circle varieties. CI: confidence interval; EIT: elicited imitation test.

DET tasks ($\beta = .079$, $p < .001$), while the other listener groups performed comparably. This finding is somewhat unexpected, given that Korean listeners' EIT scores were lower than those of other listener groups, as shown in Table 2.

We also found a shared L1 boosting effect and an outer-circle (Indian accent) weakening effect. That is, test-takers performed significantly better when listening to their own L1 accents as compared with other expanding-circle accents ($\beta = .016$, $p = .044$); in contrast, when they listen to Indian accents, their performance dropped ($\beta = -.041$, $p = .035$). In addition, participants demonstrated a positive score increase when listening to inner-circle accents ($\beta = .019$) as compared with listening to other expanding-circle varieties, but this result did not reach statistical significance ($p = .416$). This result can be further clarified in Table 4. When inner-circle varieties were used as a reference group, listeners' DET listening test scores dropped, $\beta = -.019$, which means that there was a slight negative trend when they listened to expanding-circle L2 varieties, but no statistical significance was achieved. In general, shared-L1 benefit effects emerged, but listeners' test scores did not differ significantly when they listened to inner-circle English accents or highly intelligible expanding-circle L2 varieties.

One trend that caught our attention in the model is the lack of differences in listeners' performance on tasks with inner-circle and shared-L1 accents. To further investigate this trend, the order of fixed effects in the model was rearranged to allow for a direct comparison between these two accent varieties. In this model, summarized in Table 4, the reference

Table 5. Summary of the linear mixed-effects model with task included.

Predictors	Model 5		
	Estimates	CI	*p*
(intercept)	−0.028	[−0.066, 0.008]	.126
EIT	0.004	[0.003, 0.004]	<.001
Listener background: Korean	0.079	[0.043, 0.115]	<.001
Listener background: Indian	−0.021	[−0.060, 0.017]	.274
Listener background: Mexican	0.034	[−0.003, 0.070]	.071
Inner-circle accent	0.019	[−0.026, 0.063]	.416
Outer-circle (Indian) accent	−0.041	[−0.079, −0.003]	.035
Shared L1 accent	0.016	[0.000, 0.032]	.044
Task	0.002	[−0.009, 0.013]	.708
Random effects			
σ^2	0.03		
τ_{00}	0.005	Listener	
	0.002	Speaker	
ICC	0.214		
N	160	Listener	
	24	Speaker	
Observations	3564		

Note: Reference group for listener background = Chinese; reference group for speaker accent = other expanding-circle varieties. CI: confidence interval; EIT: elicited imitation test.

group for accent comparisons was inner-circle varieties. The model confirmed our earlier observation and showed a lack of significant difference between listeners' scores on tasks with inner-circle and shared-L1 accents. This suggests that listeners performed equally well on tasks that included accents like American and British English and accents of speakers with shared L1 backgrounds as the listeners. In sum, the listener participants in this study performed comparably with their own highly intelligible accent, inner-circle accents, and other highly intelligible L2 accents.

One caveat is that the lack of significance in inner-circle varieties with other expanding-circle accents could result from a fairly large standard error (95% CI for beta ranges between −0.026 and 0.063). This indicates that there is a wide variability across listeners in terms of how their performance on inner-circle accents compares with that of other expanding accents. Also, the lack of a significant listener–speaker background interaction suggests that the shared L1 boosting effect is consistent across listener backgrounds.

The effect of task on DET listening

The second research question further investigated the role of different accents in test-taker's performance on the two listening task types of the DET. Table 5 summarizes the model with task included as an additional fixed factor. When task was added as an independent

variable in the linear mixed-effects model, task did not emerge as a significant predictor of standardized task scores ($\beta = .002$, $p = .708$). The overall fit of the model also worsened with slight increases of Akaike information criterion (AIC) and Bayesian information criterion (BIC) values from −2337.8 to −2336 and from −2269.9 to −2261.8, respectively. In addition, the explanatory power of the model did not improve either (marginal $R^2 = .116$, conditional $R^2 = .305$). This means that test-takers' scores are generalizable across DET tasks, adding supportive evidence for the generalization inference for DET test scores.

Discussion

The effects of shared versus different English accents

A major research aim of the current project was to examine the impact of different or shared-L1 accent varieties on listeners' performances on the DET listening tests. We included two inner-circle English varieties (American and British) as well as one outer-circle (Indian) and three expanding-circle accents (Chinese, Mexican, and Korean English) as speakers. Listeners from the same four noninner circle backgrounds as the speakers who recorded the test prompts listened to three different sets of accents when they took DET listening tests: (1) L1 inner-circle accent, (2) shared-L1 accent, and (3) nonshared L2 accent. Our findings showed that there was a shared L1 boosting effect. In other words, listeners generally performed better when listening to their own L1 accents as compared with listening to other unfamiliar L2 accents. In addition, there was no statistically significant difference in listeners' test scores on DET tasks between shared-L1 accents and inner-circle English accents and also no difference between inner-circle and expanding-circle accent contexts. This means that with reference to the expanding-circle accents, listeners' test scores were more positive when they heard their own shared-L1 accents, but when compared with those from the inner-circle accents (American and British English), these scores were analogous. However, listeners' performances decreased significantly when they listened to an Indian accent, even though those speakers were highly intelligible.

This finding about the shared-L1 boosting effect directly supports Bent and Bradlow's (2003) interlanguage speech intelligibility benefit and Best's (1995) claim about benefits of listener familiarity with the phonological patterns of the speaker's accent, as listeners showed some advantages of shared-L1 and/or accent familiarity in their performance. Previous research reported some inconsistent patterns regarding such shared-L1 benefits (e.g., Kang et al., 2018; Shin et al., 2021), but the linear mixed-effect models fit on the data in the current investigation generated a significant positive effect with no significant listener–speaker background interaction. In fact, the observed shared-L1 benefit specifically makes sense because the current DET listening tests included dictation and yes/no word-identification tasks, which are a common method for general intelligibility measures (Kang et al., 2020). Most interlanguage speech intelligibility benefit research (e.g., Aryadoust & Luo, 2022; Bent and Bradlow, 2003; Nishizawa, 2023) tends to employ this type of construct operationalization (i.e., transcription or dictation), which could lead to parallel findings, to some extent. On the contrary, other listening assessment studies often use passage-based listening comprehension tests, which could require different

aspects of cognitive processes, and therefore, subsequent outcomes can be more compounded. In addition, more recent research has employed electroencephalography (EEG) to investigate an interlanguage benefit in syntactic processing (Gosselin et al., 2022). As Harding (2012) claimed, the observation of shared-L1 effects can be influenced by various interfering variables, and one of them would be test-taker productions that arise from the dictates of different item types.

At the same time, the four L1 groups (Chinese, Korean, Spanish, and Indian) selected for this study have already been investigated through other previous studies that detected some shared-L1 effects. For example, Chinese shared-L1 benefit effects were revealed in Harding (2012), Dai and Roever (2019), and Shin et al. (2021), Spanish-shared L1 effects were demonstrated in Major et al. (2002), Korean shared-L1 effects were yielded in Shin et al. (2021), and Kang et al. (2019) found Indian shared-L1 effects. Even though findings of this prior research have been mixed, it is possible that those four selected L1 groups have some probability of holding shared-L1 effects, especially when speakers are highly intelligible. Further research could establish whether the result in the current study is was due to the characteristics of the L1 groups or, rather, to the nature of the DET listening tasks.

Unsurprisingly, listeners performed correspondingly well with inner-circle L1 accents (i.e., American and British). There was no significant difference in their test scores between their shared-L1 accents and inner-circle L1 accents. Listeners' test scores with American and British accents were somewhat higher than those with nonshared L2 accents, although this difference was not statistically significant in the current analysis. Indeed, this result aligns with those in many other studies (e.g., Kang et al., 2018; Major et al., 2002), in which English learners' high-stakes listening test scores were much better when they took the tests with an American accent. As Kang et al. (2019) argue, listeners' positive performances on listening tasks spoken with the inner-circle accents might still be related to listeners' familiarity with particular accents (Gass & Varonis, 1984). In other words, the participants in the study might have more exposure to these two prestigious accents, while their exposure to other varieties could be more limited. Repeated exposure to a certain accent can lead to accent familiarity, which can aid listening comprehension (Bradlow & Bent, 2008; Weber et al., 2011). We speculate that this is one of the reasons why listening task scores with the inner-circle accent were comparable to those with shared L1 accents namely because listeners were familiar with these two types of accents: either their own shared L1 or prestigious L1s.

However, there was also a lack of significant difference between inner-circle and other expanding-circle accents (in Table 4), even though listeners' test scores were slightly lower with the expanding-circle accents. This result should be interpreted carefully because there was a fairly large standard error with a wide variability across listeners in the current study, particularly when they completed the tasks with inner-circle accents in comparison with other expanding accents. Nevertheless, this finding provides evidence supporting the inclusion of different accent varieties in the listening component of high-stakes proficiency tests (Taylor, 2006). It also supports Kang et al.'s (2019) argument that as long as speakers are highly comprehensible and intelligible, they can be incorporated into high-stakes listening comprehension tests, as test-takers' listening test scores do not differ across speakers who were carefully and rigorously selected.

In sum, in the simulated DET listening tests with a variety of English accents introduced, the listener's shared-L1 accent showed positive effects. No statistical difference was found in listeners' test scores between the listener's shared-L1 accent and prestigious English accents. Also, no dissimilarity was uncovered between inner-circle models and other nonshared expanding-circle varieties. Although the outer-circle (Indian) accent indicated a negative effect, these findings are promising because they suggest the potential of incorporating other varieties of highly intelligible non-Anglophone speakers in high-stakes listening tests, which could address the issues of fairness and ecological validity (Hamp-Lyons, & Davies, 2008). More specifically, it could inform future test development by possibly introducing an option to test-takers, that is, either inner-circle L1s (e.g., American accent) or a highly intelligible shared-L1 accent. Although the interplay between listeners' experience with particular accents and their test performance can be a complicated process and needs further validation, such an active implementation in L2 assessment can offer more ecologically valid opportunities and positive washback effects to test-takers.

The effects of different accents on different listening tasks

We further examined the effects of different English accents on listeners' performance in two different task types (i.e., "yes/no" vocabulary and dictation), given that the type of input that listeners receive could affect their listening test scores (Wagner, 2014). We found no task difference effect between dictation and yes/no vocabulary tasks. As seen in Table 5, when task was added to the linear mixed-effects model, it did not emerge as a significant predictor of standardized task scores ($p = .708$). This means that test-takers' test scores did not differ in both dictation and yes/no vocabulary tasks. This finding could support Pisoni and Luce's (1987) view that both dictation and yes/no vocabulary identification tasks can be treated as a speaker intelligibility measure, because both require speech processing at the phonemic level, knowledge of sound co-occurrences, and working memory involvement. Again, the DET dictation task is similar to the speaker intelligibility measure commonly used in research studies, in which listeners try to correctly transcribe words and sentences they hear (Kang et al., 2020). Since both dictation and yes/no vocabulary tasks in the DET count on phonemic speech processing (Cohen, 1994), the most influential characteristics for listeners is likely speakers' intelligibility or clarity of speech. This is perhaps the reason that there was no task effect detected in the current study.

This outcome is encouraging and useful for future test item development. Given that there was no interplay between listening tasks and accent varieties, particularly for word recognition (yes/no vocabulary) and sentence dictation items, we can suggest that a proficiency test like DET can incorporate different English accent varieties in the listening tests interchangeably, if and when needed.

Conclusion and limitations

This study aimed to provide guidance to help test practitioners understand the impact of different varieties of English on listener performances in high-stakes tests. To

summarize, when listeners took the DET listening test from the sentence dictations and word recognitions recorded by highly intelligible speakers who carried their own L1 speech characteristics, there was a consistent shared-L1 benefit effect. Both shared-L1 and inner-circle L1 accents were also comparable in terms of listeners' listening performance scores. In addition, when listeners listened to other highly intelligible expanding-circle varieties (except for Indian accents), their listening performance scores did not differ significantly. Since there was no task effect, this result was consistent across the task types.

An attempt to use highly intelligible WE speakers can reduce the potential for unequal construct representation across groups and can have greater face validity with stakeholders (Harding, 2012). Importantly, the use of only one variety of English in a listening test may underrepresent the measured construct, particularly when they need to communicate with learners from other L1 backgrounds in target language use settings (e.g., university campuses with large numbers of international students; Ockey & French, 2016). In this study, we explored the idea of opportunity fairness by supporting accent varieties. This effort could promote positive washback effects in language classrooms (Ockey & French, 2016) and inform material development in the test preparation industry. Our findings demonstrated that listeners' own accents were beneficial to their listening test performance.

Our findings may not be generalizable to other GE communities due to the limited number of GE speakers chosen from within each circle. Also, speakers and listeners selected in this study may not represent the varieties of English at issue. In particular, even though we framed our arguments from the context of GE, this study included only one GE outer-circle variety, that is, Indian accents. A wider range of GE speakers from each of the concentric circles should be included in future research. Furthermore, it should be noted that our recommendation about the inclusion of GE varieties is restricted to highly intelligible speakers determined by a systematic speaker-screen-in method. The current project selected 24 speakers out of 77 potential speaker participants. It took 4–5 months to identify suitable speakers to meet the standards (e.g., score at least four out of five or higher on accentedness and intelligibility and at least 90% on intelligibility). An additional or follow-up study could be conducted to explore methodological protocols to identify intelligible speaker thresholds and develop the speaker selection mechanism.

Moreover, some parts of the current results could be attributable to the familiarity of the test type from the group of Korean listeners, as seen from their score pattern differences in EIT and DET scores. Future research could examine the relationships among test type familiarity, accent varieties, and listener background. Finally, we made an exploratory and provisional recommendation about integrating other accent varieties into listening tests that have similar task types as those investigated in our study measuring nonlistening comprehension constructs by suggesting that test-takers could choose their preferred accent model. Further research should investigate test-takers' reactions to and attitudes toward a greater diversity of accents on the listening section of high-stakes tests and establish the criterion validity of this test practice with other conventional listening tests.

Author contributions

Okim Kang: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Resources; Software; Supervision; Validation; Visualization; Writing—original draft; Writing—review & editing.
Xun Yan: Conceptualization; Formal analysis; Funding acquisition; Investigation; Methodology; Resources; Software; Supervision; Validation; Writing—review & editing.
Maria Kostromitina: Data curation; Formal analysis; Investigation; Methodology; Project administration; Software; Validation; Visualization; Writing—original draft; Writing—review & editing.
Ron Thomson: Conceptualization; Funding acquisition; Writing—review & editing.
Talia Isaacs: Conceptualization; Funding acquisition; Writing—review & editing.

Declaration of conflicting interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Maria Kostromitina (3rd author) currently holds a position as an assessment specialist at Duolingo English Test. However, the study was conducted, completed, and accepted for Language Testing before Maria Kostromitina started her role at Duolingo. The researchers did not receive any special access to Duolingo resources. Xun Yan (2nd author) and Talia Isaacs (5th author) currently serve on the Editorial Board of *Language Testing* as Co-Editors. An earlier version of this manuscript was submitted to *Language Testing* and was initially reviewed before Talia Isaacs was appointed to the role of Co-Editor of the journal in January 2023 and prior to Xun Yan's appointment to the co-editorship in 2024. Former Co-Editor Paula Winke conducted all editorial processing for this manuscript. The remaining authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The study was conducted with support from the Competitive Duolingo Research Grant Program. The views expressed are solely those of the authors.

ORCID iDs

Okim Kang https://orcid.org/0000-0002-7721-5283
Xun Yan https://orcid.org/0000-0002-4544-4938
Talia Isaacs https://orcid.org/0000-0003-4302-3379

Supplemental material

Supplemental material for this article is available online.

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

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

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior research methods*, *52*, 388–407. https://doi.org/10.3758/s13428-019-01237-x

Aryadoust, V., & Luo, L. (2022). The typology of second language listening constructs: A systematic review. *Language Testing*, *40*, 375–409. https://doi.org/10.1177/02655322221126604

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beeckmans, R., Eyckmans, J., Janssens, V., & Dufranne, M. (2001). Examining the yes/no vocabulary test: Some methodological issues in theory and practice. *Language Testing*, 18, 235–274. https://doi.org/10.1177/026553220101800301

Bent, T., & Bradlow, A. R. (2003). The interlanguage speech intelligibility benefit. *Journal of the Acoustical Society of America*, *114*(3), 1600–1610. https://doi.org/10.1121/1.1603234

Best, C. T. (1995). A direct realist perspective on cross-language speech perception. In W. Strange (Ed.), *peech perception and linguistic experience: Theoretical and methodological issues in cross-language speech research* (pp. 167–200). York Press.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729. https://doi.org/10.1016/j.cognition.2007.04.005

Brown, J. D. (2014). The future of world Englishes in language testing. *Language Assessment Quarterly*, *11*(1), 5–26. https://doi.org/10.1080/15434303.2013.869817

Brown, J. D., & Trace, J. (2018). Connected-speech dictations for testing listening. In H. Kayi-Aydar & J. Reinhardt (Eds.), *Language learning and language teaching* (pp. 45–63). John Benjamins. https://doi.org/10.1075/lllt.50.04bro

Buck, G. (1994). The appropriacy of psychometric measurement models for testing second language listening comprehension. *Language Testing*, *11*(2), 145–170. https://doi.org/10.1177/026553229401100204

Buck, G. (2001). *Assessing listening*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732959

Cai, H. (2013). Partial dictation as a measure of EFL listening proficiency: Evidence from confirmatory factor analysis. *Language Testing*, *30*(2), 177–199. https://doi.org/10.1177/0265532212456833

Canagarajah, S. (2006). Changing communicative needs, revised assessment objectives: Testing English as an international language. *Language Assessment Quarterly: An International Journal*, *3*(3), 229–242. https://doi.org/10.1207/s15434311laq0303_1

Carey, M. D., Mannell, R. H., & Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing*, *28*(2), 201–219. https://doi.org/10.1177/0265532210393704

Carey, M. D., & Szocs, S. (2023). Revisiting raters' accent familiarity in speaking tests: Evidence that presentation mode interacts with accent familiarity to variably affect comprehensibility ratings. *Language Testing*, *41*(2), 290–315. https://doi.org/10.1177/02655322231200808

Cohen, A. D. (1994). Verbal reports on learning strategies. *TESOL Quarterly*, *28*(4), 678–682. https://doi.org/10.2307/3587555

Dai, D. W., & Roever, C. (2019). Including L2-English varieties in listening tests for adolescent ESL learners: L1 effects and learner perceptions. *Language Assessment Quarterly*, *16*(1), 64–86. https://doi.org/10.1080/15434303.2019.1601198

Davies, A. (2009). Assessing world Englishes. *Annual Review of Applied Linguisics*, *29*, 80–89. https://doi.org/10.1017/S0267190509090072

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, *19*(1), 1–16. https://doi.org/10.1017/S0272263197001010

Eberhard, D. M., Simmons, G. F., & Fenning, C. D., (Eds.). (2022). *Ethnologue: Languages of the world* (25th ed.). SIL International.

Field, J. (2013). Cognitive validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening* (pp. 77–151). Cambridge University Press.

Flowerdew, J. (1994). Research of relevance to second language lecture comprehension: An overview. In J. Flowerdew (Ed.), *Academic listening: Research perspectives* (pp. 7–29). Cambridge University Press. https://doi.org/10.1017/CBO9781139524612.004

Fox, J. (2003). Effect displays in R for generalised linear models. *Journal of Statistical Software*, *8*(15), 1–27. https://www.jstatsoft.org/article/view/v008i15

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). https://doi.org/10.18637/jss.v008.i15

Galloway, N., & Rose, H. (2015). *Introducing global Englishes*. Routledge. https://doi.org/10.4324/9781315734347

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

Gosselin, L., Martin, C. D., Gonzalez Martin, A., & Caffarra, S. (2022). When a nonnative accent lets you spot all the errors: Examining the syntactic interlanguage benefit. *Journal of Cognitive Neuroscience*, *34*(9), 1650–1669. https://doi.org/10.1162/jocn_a_01886

Hamid, M. O., Hardy, I., & Reyes, V. (2019). Test-takers' perspectives on a global test of English: Questions of fairness, justice, and validity. *Language Testing in Asia*, *9*, 1–20. https://doi.org/10.1186/s40468-019-0092-9

Hamp-Lyons, L., & Davies, A. (2008). The English of English tests: Bias revisited. *World Englishes*, *27*(1), 27–39. https://doi.org/10.1111/j.1467-971X.2008.00534.x

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

Harding, L. & McNamara, T.F. (2018). Language assessment: The challenge of ELF. In J. Jenkins, M. J. Dewey, & W. Baker (Eds.), *Routledge Handbook of English as a Lingua Franca*. Routledge. https://doi.org/10.4324/9781315717173-46

Harrington, M., & Carey, M. (2009). The on-line yes/no test as a placement tool. *System*, *37*(4), 614–626. https://doi.org/10.1016/j.system.2009.09.006

Harsch, C., & Hartig, J. (2016). Comparing c-tests and yes/no vocabulary size tests as predictors of receptive language skills. *Language Testing*, *33*(4), 555–575. https://doi.org/10.1177/0265532215594642

Isaacs, T., & Rose, H. (2022). Redressing the balance in the native speaker debate: Assessment standards, standard language, and exposing double standards. *TESOL Quarterly*, *56*, 401–412. https://doi.org/10.1002/tesq.3041

Jia, C., & Hew, K. F. T. (2019). Supporting lower-level processes in EFL listening: The effect on learners' listening proficiency of a dictation program supported by a mobile instant messaging app. *Computer Assisted Language Learning*, *35*, 141–168. https://doi.org/10.1080/09588221.2019.167146

Jenkins, J., Cogo, A., & Dewey, M. (2011). Review of developments in research into English as a lingua franca. *Language teaching*, *44*(3), 281–315. https://doi.org/10.1017/S0261444811000115

Kachru, B. B. (1985). Standard, codification and sociolinguistic realism: The English language in the outer circle. In R. Quirk & H. Widdowson (Eds.), *English in the world: Teaching and learning the language and literatures* (pp. 11–30). Cambridge University Press.

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

Kang, O. (2015). Students' perceptions of pronunciation instruction in the three circles of World Englishes. *TESOL Journal*, *6*(1), 59–80. https://doi.org/10.1002/tesj.146

Kang, O., Ahn, H., Yaw, K., & Chung, S. (2021). *Investigation of relationship among learner background, linguistic progression, and score gain on IELTS*. The IELTS Research Report Series. https://www.ielts.org/-/media/research-reports/ielts-rr_2021-1_kang-et-al.ashx

Kang, O., & Moran, M. (2019). Enhancing communication between ITAs and US undergraduate students. In S.D. Looney & S. Bhalla (Eds.), *A transdisciplinary approach to international teaching assistants: Perspectives from applied linguistics* (pp. 82–89). Multilingual Matters.

Kang, O., Moran, M., Ahn, H., & Park, S. (2020). Proficiency as a mediating variable of intelligibility for different varieties of accents. *Studies in Second Language Acquisition*, *42*(2), 471–487. https://doi.org/10.1017/S0272263119000536

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. https://doi.org/10.1111/lang.12270

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*(1), 115–146. https://doi.org/10.1111/lang.12270

Kim, Y. S., Tracy-Ventura, N., & Jung, Y. (2016). A measure of proficiency or short-term memory? Validation of an elicited imitation test for SLA research. *The Modern Language Journal*, *100*(3), 655–673. https://doi.org/10.1111/modl.12346

Kostromitina, M., & Plonsky, L. (2022). Elicited imitation tasks as a measure of L2 proficiency: A meta-analysis. *Studies in Second Language Acquisition*, *44*(3), 886–911. https://doi.org/10.1017/S0272263121000395

Kraljic, T., Brennan, S. E., & Samuel, A. G. (2008). Accommodating variation: Dialects, idiolects, and speech processing. *Cognition*, *107*(1), 54–81. https://doi.org/10.1016/j.cognition.2007.07.013

Kremmel, B., & Schmitt, N. (2016). Interpreting vocabulary test scores: What do various Item formats tell us about learners' ability to employ words? *Language Assessment Quarterly*, *13*(4), 377–392. https://doi.org/10.1080/15434303.2016.1237516

Kunnan, A. (2008). Towards a model of test evaluation: Using the test fairness and test context frameworks. In L. Taylor & C. J. Weir (Eds.), *Multilingualism and assessment: Achieving transparency, assuring quality, sustaining diversity* (pp. 229–251). Cambridge University Press.

Kunnan, A. (2014). Fairness and justice in language assessment. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1–17). John Wiley & Sons. https://doi.org/10.1002/9781118411360

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

LaFlair, G. T., & Settles, B. (2020). *Duolingo English Test: Technical manual* [Duolingo research report]. https://englishtest.duolingo.com/research

Lagrou, E., Hartsuiker, R., & Duyck, W. (2013). The influence of sentence context and accented speech on lexical access in second-language auditory word recognition. *Bilingualism: Language and Cognition*, *16*(3), 508–517. https://doi.org/10.1017/S1366728912000508

Lang, M., & R Core Team. (2020). *backports: Reimplementations of functions introduced since R-3.0.0* (R package version 1.1.9). https://CRAN.R-project.org/package=backports

Llurda, E. (2004). Non-native-speaker teachers and English as an International Language. *International Journal of Applied Linguistics*, *14*(3), 314–323. https://doi.org/10.1111/j.1473-4192.2004.00061.x

Long, J. A. (2019). *Interactions: Comprehensive, user-friendly toolkit for probing interactions* (R package version 1.1.0). https://cran.r-project.org/package=interactions

Lüdecke, D. (2021). *sjPlot: Data visualization for statistics in social science* (R package version 2.8.10). https://CRAN.R-project.org/package=sjPlot

Major, R. C., Fitzmaurice, S. F., Bunta, F., & Balasubramanian, C. (2002). The effects of nonnative accents on listening comprehension. *TESOL Quarterly*, *36*(2), 173–190. https://doi.org/10.2307/3588288

Matthews, J., & Cheng, J. (2015). Recognition of high frequency words from speech as a predictor of L2 listening comprehension. *System*, *52*, 1–13. https://doi.org/10.1016/j.system.2015.03.006

McArthur, T. (1998). *The English languages*. Cambridge University Press. https://doi.org/10.1017/9780511621048

Milton, J., Wade, J., & Hopkins, N. (2010). Aural word recognition and oral competence in English as a foreign language. In R. Chacón-Beltrán, C. Abello-Contesse, & M. del Mar Torreblanca-López (Eds.), *Insights into non-native vocabulary teaching and learning* (pp. 83–98). Multilingual Matters. https://doi.org/10.21832/9781847692900-007

Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition*, *28*(1), 111–131. https://doi.org/10.1017/S0272263106060049

Nation, I. S. P., & Newton, J. (2009). *Teaching ESL/EFL listening and speaking*. Routledge. https://doi.org/10.4324/9780203886489

Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology*, *18*(1), 62–85. https://doi.org/10.1177/0261927X99018001005

Nishizawa, H. (2023). Construct validity and fairness of an operational listening test with World Englishes. *Language Testing*, *40*(3), 493–520. https://doi.org/10.1177/02655322221137869

Ockey, G. J., & French, R. (2016). From one to multiple accents on a test of L2 listening comprehension. *Applied Linguistics*, *37*(5), 693–715. https://doi.org/10.1093/applin/amu060

Ortmeyer, C., & Boyle, J. P. (1985). The effect of accent differences on comprehension. *RELC Journal*, *16*(2), 48–53. https://doi.org/10.1177/003368828501600205

Pellicer-Sánchez, A., & Schmitt, N. (2012). Scoring yes–no vocabulary tests: Reaction time vs. nonword approaches . *Language Testing*, *29*(4), 489–509. https://doi.org/10.1177/0265532212438053

Pisoni, D. B., & Luce, P. A. (1987). Acoustic-phonetic representations in word recognition. *Cognition*, *25*(1-2), 21–52. https://doi.org/10.1016/0010-0277(87)90003-5

Revelle, W. (2020). *psych: Procedures for personality and psychological research* (R package version 2.0.9). https://CRAN.R-project.org/package=psych

Rose, H., McKinley, J., & Galloway, N. (2021). Global Englishes and language teaching: A review of pedagogical research. *Language Teaching*, *54*(2), 157–189. https://doi.org/10.1017/S0261444820000518

Ross, S., & Langille, J. (1997). Negotiated discourse and interlanguage accent effects on a second language listening test. In G. Brindley & G. Wigglesworth (Eds.), *Access: Issues in language test design and delivery* (pp. 87–116). National Centre for English Language Teaching and Research, Macquarie University.

Saito, K., Tran, M., Suzukida, Y., Sun, H., Magne, V., & Ilkan, M. (2019). How do second language listeners perceive the comprehensibility of foreign-accented speech? Roles of first language profiles, second language proficiency, age, experience, familiarity, and metacognition. *Studies in Second Language Acquisition*, *41*(5), 1133–1149. https://doi.org/10.1017/S0272263119000226

Shin, S. Y., Lee, S., & Lidster, R. (2021). Examining the effects of different English speech varieties on an L2 academic listening comprehension test at the item level. *Language Testing*, *38*(4), 580–601. https://doi.org/10.1177/0265532220910109

Siegel, J., & Siegel, A. (2015). Getting to the bottom of L2 listening instruction: Making a case for bottom-up activities. *Studies in Second Language Learning and Teaching*, *5*(4), 637–662. https://doi.org/10.14746/ssllt.2015.5.4.6

Smith, L. E., & Bisazza, J. A. (1982). The comprehensibility of three varieties of English for college students in seven countries. *Language Learning*, *32*(2), 259–269. https://doi.org/10.1111/j.1467-1770.1982.tb01307.x

Stibbard, R. M., & Lee, J. I. (2006). Evidence against the mismatched interlanguage speech intelligibility benefit hypothesis. *The Journal of the Acoustical Society of America*, *120*(1), 433–442. https://doi.org/10.1121/1.2203683

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/cci076

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.2011.02.004

Wagner, E. (2014). Using unscripted spoken texts in the teaching of second language listening. *TESOL Journal*, *5*(2), 288–311. https://doi.org/10.1002/tesj.92

Weber, A., Broersma, M., & Aoyagi, M. (2011). Spoken-word recognition in foreign accented speech by L2 listeners. *Journal of Phonetics*, *39*(4), 479–491. https://doi.org/10.1016/j.wocn.2011.02.006

Weir, C. (1993). *Understanding and designing language tests*. Longman.

Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, *40*(1), 1–29. https://doi.org/10.18637/jss.v040.i01

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer. https://doi.org/10.1007/978-3-319-24277-4

Wickham, H., & Bryan, J. (2019). *readxl: Read excel files* (R package version 1.3.1). https://CRAN.R-project.org/package=readxl

Wickham, H, Averick, M, Bryan, J, Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Petersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., . . .Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), Article 1686. https://doi.org/10.21105/joss.01686

Yan, X. (2020). Unpacking the relationship between formulaic sequences and speech fluency on elicited imitation tasks: Proficiency level, sentence length, and fluency dimensions. *TESOL Quarterly*, *54*(2), 460–487. https://doi.org/10.1002/tesq.556

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. https://doi.org/10.1177/0265532215594643

Yule, G., Wetzel, S., & Kennedy, L. (1990). Listening perception accuracy of ESL learners as a variable function of speaker L1. *TESOL Quarterly*, *24*, 519–523. https://doi.org/10.2307/3587244

Zhang, Q. (2022). Impacts of World Englishes on local standardized language proficiency testing in the Expanding Circle: A study on the College English Test (CET) in China. *English Today*, *38*(4), 254–270. https://doi.org/10.1017/S0266078421000158

Zhang, X., Liu, J., & Ai, H. (2019). Pseudowords and guessing in the Yes/No format vocabulary test. *Language Testing*, *37*(1), 6–30. https://doi.org/10.1177/0265532219862265

Appendix 1. Summary of model fit in research questions 1 and 2.

Predictors	Model 0			Model 1			Model 2			Model 3			Model 4			Model 5		
	Estimates	*CI*	*p*	Estimates	*CI*	*p*	Estimates	*CI*	*p*	Estimates	*CI*	*p*	Estimates	*CI*	*p*	Estimates	*CI*	*p*
(intercept)	0.002	[−0.014, 0.019]	.771	0.001	[−0.024,0.026]	.91	0.001	[−0.022, 0.025]	.920	−0.01	[−0.039, 0.019]	.155	−0.028	[−0.065, 0.009]	.136	**−0.029**	**[−0.067, 0.008]**	**.126**
EIT							0.003	[0.003, 0.004]	<.001	0.003	[0.002, 0.004]	<.001	0.004	[0.003, 0.004]	<.001	**0.004**	**[0.003, 0.004]**	**<.001**
Listener background: Korean										0.078	[0.042, 0.114]	<.001	0.079	[0.043, 0.115]	<.001	**0.079**	**[0.043, 0.115]**	**<.001**
Listener background: Indian										−0.012	[−0.049, 0.026]	.542	−0.021	[−0.060, 0.017]	.274	−0.021	[−0.060, 0.017]	0.274
Listener background: Mexican										0.035	[−0.002, 0.071]	.063	0.034	[−0.003, 0.070]	.071	**0.034**	**[−0.003, 0.070]**	**.071**
Inner-circle accent													0.019	[−0.026, 0.063]	.416	0.019	[−0.026, 0.063]	.416
Outer circle accent													−0.041	[−0.079, −0.003]	.035	**−0.041**	**[−0.079, −0.003]**	**0.035**
Share L1 accent													0.016	[0.000, 0.032]	.044	**0.016**	**[0.000, 0.032]**	**0.044**
Task																0.002	[−0.009, 0.013]	0.708
Random effects																		
q2	0.03			0.03			0.03			0.03			0.03			0.03		
t00	0.01	Listener		0.01	Listener		0.006	Listener		0.005	Listener		0.005	Listener		0.005	Listener	
				0.002	Speaker		0.002	Speaker		0.002	Speaker		0.002	Speaker		0.002	Speaker	
ICC	0.244			0.296			0.236			0.213			0.214			0.214		
N	160	Listener		160	Listener		160	Listener		160	Listener		160	Listener		160	Listener	
				24	Speaker		24	Speaker		24	Speaker		24	Speaker		24	Speaker	
Observations	3564			3564			3564			3564			3564			3564		
Deviance	−2098.76			−2257.60			−2312.11			−2339.02			−2359.8			−2360		
log-Likelihood	1049.4			1128.8			1156.1			1169.5			1179.9			1180		
Chi-square difference				158.85			54.52			26.959			20.759			0.143		
p				<.001			<.001			<.001			<.001			.705		
AIC	−2092.8			−2249.6			−2302.1			−2323.1			−2337.8			−2336		
BIC	−2074.2			−2224.9			−2271.2			−2273.7			−2269.9			−2261.8		

Reference group for listener background = Chinese; reference group for speaker accent = other expanding-circle varieties. AIC: Akaike information criterion; BIC: Bayesian information criterion; CI: confidence interval; EIT: elicited imitation test;.