Including L2-English Varieties in Listening Tests for Adolescent ESL Learners: L1 Effects and Learner Perceptions

David Wei Dai, UCL Institute of Education, University College London,

david.dai@ucl.ac.uk

https://orcid.org/0000-0002-3575-131X

Carsten Roever, School of Languages and Linguistics, University of Melbourne,

carsten@unimelb.edu.au

https://orcid.org/0000-0002-1055-6331

This is an author-produced PDF of an article accepted for publication in *Language Assessment Quarterly* following peer review. The page numbers of this version are formatted to match the ones of the definitive publisher-authenticated version, which can be found at:
https://doi.org/10.1080/15434303.2019.1601198

To cite (APA): 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

Abstract

Because English is widely used as a lingua franca, language testers have started to consider the introduction of non-native accents into English listening tests. This study investigates how accents influence test-takers' performance, and also elicits test-takers' subjective perception of accents. Eighty adolescent L1-Mandarin test takers were divided into four groups of equal proficiency, with each group listening to one accented version of the same English listening test. The test input was delivered in Australian, Spanish, Vietnamese, and Mandarin-accented English varieties with tasks measuring lexical and propositional comprehension and consisting of picture selection, true/false, and gap completion items. Test-takers' perceptions of accent familiarity, comprehensibility, and their attitudes were also measured. Results indicate that the test takers who received Mandarin-accented input performed best, lending support to a shared-L1 effect, with the strongest advantage for lexical comprehension. No significant difference was observed in test scores among the groups exposed to non-Mandarin accents. Findings also reveal that the type of accent was not significantly related to test-takers' attitude toward it. The central implication from this study is that there is potential for the inclusion of non-native accents into listening tests for adolescent learners if the shared-L1 effect can be addressed.

Introduction

Because English is being used globally in multicultural contexts among not only L1-English speakers but also L2-English speakers (Seargeant, 2012) and the number of L2-English speakers far exceeds that of L1- English speakers, there is a need for language testers to consider including L2-English varieties into high- stakes English tests for better construct representation[1]. Consequently, a growing number of English as a Lingua Franca (ELF) researchers have argued that English tests should start assessing multidialectal competence instead of using dominant English varieties as their benchmark (e.g., Brown, 2014; Jenkins & Leung, 2013; Seidlhofer, 2011; Taylor & Geranpayeh, 2011) with multidialectal competence encompassing the comprehension of standard native English, regional English, ethnic English, and non-native English accents. While there has been some research on the feasibility of diversifying traditional listening constructs that only use standard English varieties (Harding, 2011; Major, Fitzmaurice, Bunta, & Balasubramanian, 2002, 2005), such attempts have not resulted in the broadening of accent types in high- stakes tests. Major international English tests, such as TOEFL and IELTS, still limit their accent input to standard native varieties even though test takers are likely to encounter speakers of non-native accents in the target language use setting (e.g., students and academics from all over the world at universities in the United States, the United Kingdom, or Australia).

[1] The dichotomy of native/non-native, L1/L2 English accents can be problematic. However, for the sake of simplification, this study still adopts such dichotomous classifications. To be more specific, native/L1 English accents refer to the accents of speakers who use English as their first language, whereas non-native/L2 English accents refer to the accents of speakers who have learned to speak English after their first language has already been established and whose English carries phonological features of their L1.

This study examines the introduction of non-native accents into listening tests for adolescent[2] learners, motivated by the rapid recent growth of language-testing work with pre-adult populations due to the internationalization of pre-adult education and increasing mobility of this population. For example, in 2014, 73,000 international secondary students were enrolled in U.S. schools (Institute of International Education, 2014) with a strong upward trend. Similarly, 18,000 international students were enrolled in schools in Australia in 2014 (Department of Education and Training, 2014). In addition to ESL settings, English is increasingly being used as a medium of instruction in EFL settings, including state schools employing bilingual instruction and private international schools (So et al., 2015).

In all these settings, tests respond to needs for diagnosis of proficiency, learner placement, tracking of progress, program evaluation, and accountability (British Council, 2016; So et al., 2015). It is important that they incorporate multidialectal listening to reflect today's ELF reality where English speakers, whether native or non-native, need to possess the competence to understand accented English varieties (Harding & McNamara, 2017).

Background

The test perspective: Accent effects in listening tests

One of the central concerns for language testers is the effect of accent on test-takers' comprehension of the listening test, because better comprehension correlates with better performance. Comprehensibility in this article is defined as "The degree of effort required by a listener to understand an utterance" (Derwing & Munro, 2015, p. 176) and is a topic that has interested researchers inside and outside second language studies (see Sidaras, Alexander, & Nygaard, 2009; Stevenage, Clarke, & McNeill, 2012 for studies in psychology; Bent & Bradlow, 2003; Stibbard & Lee, 2006; Weber, Broersma, & Aoyagi, 2011 for studies in phonetics). Within the field of second language studies, Gass and Varonis (1984) published one of the first studies that quantitatively examined the influence of non-native accent on listener comprehension. The variable under investigation was familiarity, which was further dissected into four factors: familiarity with the topic, familiarity with non-native accents in general, familiarity with a particular non-native accent, and familiarity with a particular speaker with a non-native accent. ANOVA findings suggested all four factors as conducive to listener comprehension, which was corroborated by other research (Lobo & Yoshida, 1982; Pica & Long, 1982, both cited in Gass & Varonis, 1984; Varonis & Gass, 1982).

Although Gass and Varonis (1984) pointed to other factors such as "fluency and social variables" (p. 85) that might also impact a listener's comprehension of accented speech, accent familiarity is by far the most investigated determiner in contemporary large-scale listening assessment research. Work in L1 accent varieties has shown that increasing accent strength and decreasing accent familiarity affect test-takers' performance detrimentally (Ockey & French, 2014). For L2 accents, a particular concern within the familiarity categories is the "shared-L1 effect," a language-testing term similar to Bent and Bradlow's (2003) "interlanguage speech intelligibility benefit," and formally defined by Harding (2011) as a phenomenon that can occur when a certain group of test takers who share the same L1 with the speaker of the test recording comprehend the listening materials more easily and, consequently, perform better on the test. While Bent and Bradlow (2003) suggest that this phenomenon might extend to other non-native accented input as the "mismatched speech intelligibility benefit," Stibbard and Lee (2006) show no such effect. A possible explanation for the shared-L1 effect is that repeated exposure to an accent leads to familiarity with that accent, which in turn aids comprehension (Adank, Evans, Stuart-Smith, & Scott, 2009; Bradlow & Bent, 2008; Stevenage et al., 2012; Weber et al., 2011).

When we look for the shared-L1 effect in the listening assessment literature, evidence supporting its existence is diffuse and contradictory. Some studies, such as Abeywickrama (2013) and Butler

[2] In this study "adolescent" is defined as between 12 and 15 years of age, which corresponds to the middle school period for students in China.

(2007), where test-takers' comprehension of shared-L1 accented speech is measured by listening tests, have found no evidence of benefits. Work by Munro, Derwing, and Morton (2006) and Nejjari, Gerritsen, Van der Haagen, and Korzilius (2012) also failed to capture such an effect when their test takers are asked to rate the comprehensibility of accented stimuli, which include test-takers' L1 accent. Other studies, such as Harding (2011), have shown the facilitative effect of L1 accent for certain L1 groups. Harding recruited 212 test takers (of which 70 were Mandarin-L1 participants and 60 were Japanese-L1 participants) and recorded a listening test with one section spoken in Australian English, one section in Mandarin-accented English, and the third section in Japanese-accented English. After administering the same test to all of his test takers and using differential item functioning (DIF) to analyze individual item performances, Harding found that Mandarin-L1 test takers were noticeably advantaged on several items with the Mandarin accent, lending support to the hypothesis of a shared-L1 advantage. However, such a clear advantage was not observed in the Japanese-L1 test-taker group. The mixed findings from Harding (2011) echo the results from Major et al. (2002), who found that a Spanish-L1 test-taker group benefited from the input in Spanish- accented English, supporting the shared-L1 effect argument. However, a Chinese-L1 group in Major et al. (2002) actually performed worse with the Chinese[3] L1 accent than other test-taker groups, a finding that contradicts the shared-L1 advantage.

It should be mentioned that both Harding (2011) and Major et al. (2002) attempted to control for speaker accentedness, with Harding adopting the speaker selection method in Munro and Derwing (1995) to ensure "general intelligibility" (Harding, 2011, p. 168) and Major et al. using listening test results and listener rating to aim for "moderate foreign accents" (Major et al., 2002, p. 179). Both Harding (2011) and Major et al. (2002) also point to the many interfering variables that can affect the observation of shared-L1 effects: Harding attributed his mixed findings to issues such as item types, sample size, and test-takers' differing linguistics knowledge. Major et al. (2002) speculated that the causes could be speakers' varying accentedness, test-takers' attitude to L1-accented speech, and the prosodic similarities between the non-native accents selected.

In summary, these mixed findings create uncertainty about whether a shared L1 can actually cause noticeable differences among different test-taker groups and are likely to be one of the reasons for the very cautious uptake of a multidialectal approach to listening assessment in language tests. The existence or strength of a shared-L1 effect is not entirely clear and thus deserves further investigation, which is one focus of the current study.

The test-taker's perspective: Test-takers' accent perception

While some applied linguists might favor the use of non-native accents in listening tests for broader construct coverage, it is also useful to understand test-takers' perceptions of accents. This can elucidate the reasons for possible differential performances with different accents and provide insight into test-takers' views of the acceptability of including non-native accents in tests of English. Although there is a paucity of research investigating the test-takers' perspective on multidialectal listening assessments, sociolinguistic studies have elicited listeners' responses to accented speech samples, mostly from the angle of accent attitude. Several studies have found that listeners of English accents show a preference for native over non- native accents, whether listeners' L1 is English (Fraser & Kelly, 2012; Hiraga, 2005; Nejjari et al., 2012) or not (Korean L1 in Kim, 2007; Japanese L1 in McKenzie, 2008 and Cantonese L1 in Zhang, 2009). The predilection for native accent varieties also does not seem exclusive to English because listeners in Hendriks, Meurs, and Groot (2015) reported partiality not only toward native English accents but also native French, German, and Spanish accents. In the specific context of English language teaching and

[3] Major et al. (2002) used the term "Chinese" while Harding (2011) used the term "Mandarin Chinese." It is not clear from either article how the Chinese accent was determined because Chinese accent is not a homogenous accent due to the vast number of Chinese dialects. In this study the term "Mandarin accent" is used to specifically refer to speakers who are proficient in Mandarin Chinese and were born and raised in China.

testing, EFL learners' favorable attitude toward inner-circle English accent varieties has also been documented (Butler, 2007; Chien, 2014; Yook & Lindemann, 2013). In a rare assessment study, Butler (2007) recruited 312 elementary school Korean-L1 students and grouped them to listen to oral materials recorded in either a Korean English accent or an American English accent, which she described as "a preferred NES (native English speaking) model in many EFL contexts" (p. 737). A comprehension test based on the oral materials was then administered to the students. After the comprehension test, students were asked to report their attitude to the two accents in an attitudinal questionnaire. Results from the comprehension test indicated that the students showed no difference in their comprehension of the oral materials whether the materials were delivered with a Korean or American accent. However, the group listening to the American accent reported a more favorable attitude toward this accent in the attitude questionnaire than the group listening to Korean accented input.

Such findings are unsurprising because accent perception studies have shown that when speakers are perceived to be non-native speakers, listeners tend to scrutinize speakers' phonetic features much more closely and sometimes even penalize them for imagined errors. This bias against non-native accents is identified in both native-English-speaker listeners (Lindemann, 2017) and non-native-English-speaker listeners (Hu & Lindemann, 2009). Although Yook and Lindemann (2013) seem to demonstrate some shared-L1 affinity because their Korean listeners rated a Korean-accented speaker more favorably when they were informed of the speaker's Korean background, closer examination tells a different story: Ratings of the Korean speaker's accent still lagged far behind any native speaker accent in favorability.

Areas for research

In light of previous research, our study addresses three main issues. First, studies on accented listening tasks in language testing have predominantly recruited adult test takers (Harding, 2011; Major et al., 2002, 2005; Ockey & French, 2014), whereas few studies focus on how younger test takers respond to multidialectal listening tests (Butler, 2007, being a rare exception). However, assessment of pre-adult learners is becoming an increasingly active field of research and practical work as evidenced by the recent increase in the availability of tests aimed at this population, such as the Cambridge English Young Learners Exams for "primary and lower-secondary level students" (University of Cambridge Local Examination Syndicate, 2016), the TOEFL Primary Test for 8- to 12-year-old children (Cho et al., 2016), the TOEFL Junior Test for learners from 11 to 15 years of age (So et al., 2015), and the British Council's APTIS for Teens targeting learners from 13 to 17 years old (British Council, 2016).

Second, there is a lack of research instruments that investigate global accent effects with a sufficient number of items. For example, the study by Ockey and French (2014) focused on only 6 items, and similarly Major et al. (2002), (2005) used 4 items per accent. In contrast, Harding (2011) used 30–40 items for each accented section in his test but analyzed test-takers' performance at the item level with DIF rather than at the global level. The listening test in this study contains a larger number of items and also includes different task types because little research that investigates whether accent effects interact with task types exists. For example, are gap fill tasks more or less affected by accented listening input than tasks requiring comprehension of propositions embedded in longer stretches of discourse?

Third, for accent perception, most previous studies used brief stimulus materials to elicit listeners' responses. For example, the speech materials in both Chien (2014) and Yook and Lindemann (2013) were 20-s preexisting samples from the Speech Accent Archive collected by George Mason University (Weinberger, 2017). Because of their brevity, these materials elicited listener responses based on momentary impressions of accents. In contrast, an extended test recording as designed in this study requires listeners to engage deeply and proactively with accented materials, providing more opportunities for them to reflect on and evaluate accent familiarity, comprehensibility, and attitude.

Research questions

This study addresses the following research questions:

RQ1a: What is the difference in the performance of four L1-homogenous adolescent student test-taker groups on a listening test recorded in four different English accent versions?

RQ1b: If an accent effect exists, does it impact performance on different listening tasks differentially?

RQ2: What is the difference in how students perceive the four accents in terms of familiarity, comprehensibility, and attitude?

RQ3: What is the relationship between listening test results and test-taker reports of familiarity, comprehensibility, and attitude?

Methodology

Participants

Fifteen-year-old students of Mandarin L1 background from the same grade in a public middle school in a large city in China were recruited as participants for the test and questionnaire. Out of this group of 253 students, 80 test takers (45 boys and 35 girls) were selected for this study because they scored highest on a 25-item preliminary British English-accented listening test that was administered to all 253 students. We selected high-scoring students to limit the effect of overall proficiency on test results, which might otherwise have obscured the effect of accent.

Prior to this study, students had studied English as a school subject in formal classroom settings for 8 years, which equates to approximately 700 hours of training. It should be noted that English education in the participants' school is highly focused on grammar and reading comprehension, with little attention given to developing listening skills. Therefore, despite 700 hours of training, their listening ability was still at the beginner to low intermediate level (roughly A2 in the Common European Framework of Reference as judged by researchers and their English teachers). Students reported little English learning experience outside school aside from finishing assignments given by their English teachers, which were mainly reading and grammar exercises. Prior to the study, researchers administered a background questionnaire to students, who reported no exposure to English outside their English classes at school. This is not uncommon in middle school students in China because students of that age are usually more interested in entertainment delivered through Chinese. The questionnaire also revealed that the only English accents accessible to students were British accents from textbook recordings and Mandarin accents from their English teachers because no teachers of non-Mandarin L1 background were employed at the school.

The 80 students were divided into four groups with between-group differences kept as minimal as possible based on their test scores in the preliminary test. To achieve this, each student was assigned an ID number with the top-scoring student given 1 and the lowest scoring student given 80. The students were then progressively assigned to four groups according to the zigzag grouping method shown in Figure 1. To illustrate, students 1–4 were assigned to Groups 1–4, Student 5 was also assigned to Group 4, Student 6 to Group 3, Student 7 to Group 2, Student 8 to Group 1, Student 9 also to Group 1, Student 10 to Group 2, etc. This forward-then-backward assignment and the assignment of adjacent students to Groups 1 and 4 ensured that the final sum of ranks for all the groups was identical.

To confirm the absence of between-group differences, a one-way analysis of variance (ANOVA) was conducted with the group as the factor and the scores from the preliminary listening test as the dependent variable. Descriptive statistics are reported in Table 1 and show that the test score means (out of 25) of the four groups were nearly identical. ANOVA also revealed no significant between-

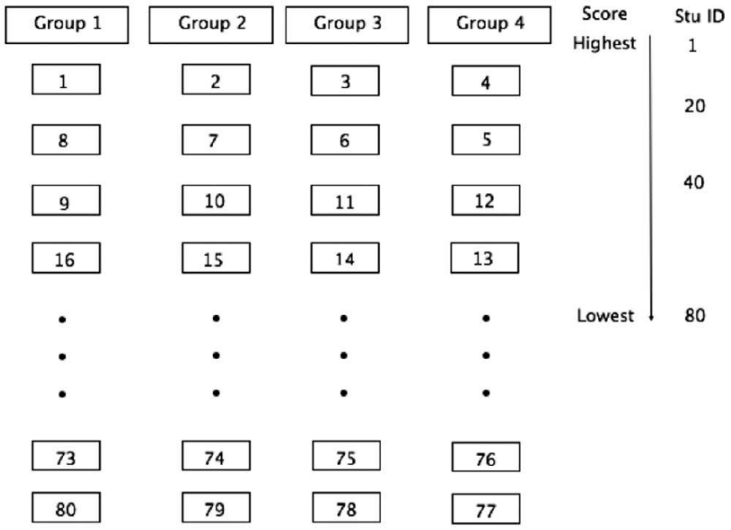


Figure 1. The grouping method to ensure equal group means.

Group	*N*	*M*	*SD*
1	20	18.00	2.71
2	20	17.95	2.76
3	20	17.90	2.65
4	20	17.90	2.91

Table 1. Descriptive statistics for the preliminary listening test.

group difference: $F_{(3,76)} = 0.006$, $p = .999$. Results indicated that the four groups were essentially identical in their English listening competence.

Instruments

The development of the final multidialectal listening test went through stages of accent selection, speaker selection, and finalizing of the test. A questionnaire was developed subsequently.

Accent selection

RQ1 requires the inclusion of four non-native English accents to record four versions of the same listening test. Because one aim of this study was to investigate the shared-L1 effect for an L1 Mandarin Chinese group, inclusion of Mandarin-accented English accent was essential. The Australian accent was chosen to focus on a less-dominant native accent to potentially offer different findings from previous studies that used British or American accents (e.g., Butler, 2007; Major et al., 2005). The two other accents were selected on the basis of their phonological features and language families. Because Mandarin is a Sino-Tibetan tonal language, a tonal Austroasiatic language (Vietnamese) was chosen to check whether tonality itself would convey an advantage. By contrast, Spanish is an Indo-European language like English but is syllable timed rather than stress timed, and the Spanish accent was therefore expected to be the one to which test takers would be the least accustomed.

Speaker selection

For the first round of speech sample recordings, 12 speakers (3 per accent) were selected. All speakers were male and between 30 and 40 years old. The reason for selecting speakers of the same gender and age group is to ensure that listeners and test takers should not be influenced by any variables other than speakers' accents. When recruiting the 12 potential speakers, the researchers relied on their experience of the four accents and selected speakers who possessed mild-to-average

accents. Speakers had resided in their L1 countries since birth until they finished tertiary education, making them users of their L1 for at least 22 years since infancy, which should lead to adequate L1 accent transfer when speaking English. Each speaker was asked to record a speech sample lasting 20s based on the script provided by researchers, and in total 12 listening samples were collected. To ensure the samples selected for the four accents had a similar degree of accent strength and accent identifiability, a Strength of Accent Scale (abbreviated as the Scale, see Appendix A) and an Accent Strength and Identification Task (abbreviated as the Task, see Appendix B) were designed. The Scale was developed on the basis of the accent scale used in Ockey and French (2014), which relied on listeners' subjective judgment to measure accent strength.

The Task was administered to a group of 35 listener judges (18 native Australian English speakers and 17 non-native English speakers). When completing the Task, they were provided the Scale to judge the strength of potential speakers' accents (see Bands 1–5 in Section II in Appendix B). Listener judges were also required to identify the possible L1 background of the speakers to ensure the final four speakers' accents were truly representative of their L1 backgrounds. The 35 listener judges were enrolled in bachelor's or master's programs in Australian universities. The non-native listener judges demonstrated advanced listening competence (at or above Band 7 in IELTS listening or equivalent) and came from various L1 backgrounds, including German, French, Estonian, Burmese, Khmer, Cantonese, and others. None of the non-native judges spoke an L1 whose accent was represented in the study. To ensure reliable accent identification, only judges who reported high familiarity with the four accents in this study and claimed to be skilled at accent judgment and identification were included. While this procedure cued the judges about what accents to expect, they still had to identify each accent among four options. In addition, because this study used mild accents, it would have been a nearly impossible task for listener judges to identify accents without any clues. The mean scores of each speaker's accent strength and identifiability are reported in Figures 2 and 3. The darker columns indicate speakers who were eventually selected for test recording (Viet 3, Spa 2, Man 3, Aus 2).

From the original 12 speakers, four speakers (1 per accent) were selected to record a version of the listening test based on the following criteria:

(1) The accent strength of the four speakers should be within a similar range, ideally, 2.0–2.5 out of 5, which represents light-to-mild accentedness on the Scale.[4]

(2) The accent identifiability of the four speakers should be >0.7, which indicates that generally 70% of the listener judges could successfully identify the speakers' first language.

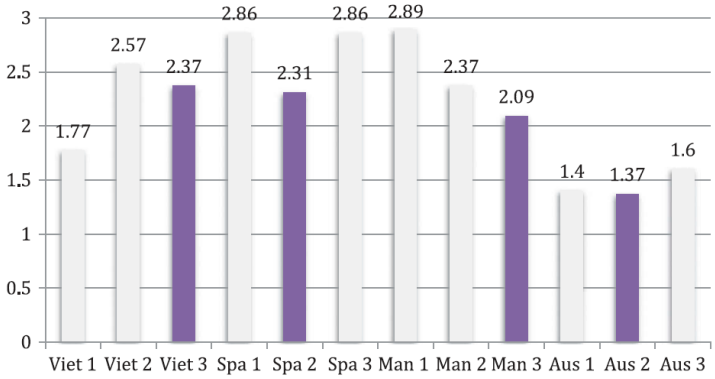


Figure 2. Mean scores of speakers' accent strength (Viet = Vietnamese, Spa = Spanish, Aus = Australian, Man = Mandarin; the bigger the mean score is the stronger the accent strength is).

[4] The accent strength of the Australian accent speaker was lower at 1.37, but this is to be expected because all three speakers for the Australian group spoke educated Australian English. In addition, a possible facilitative effect of a native accent would not be detectable if that native accent was strongly impacted by dialectal features.

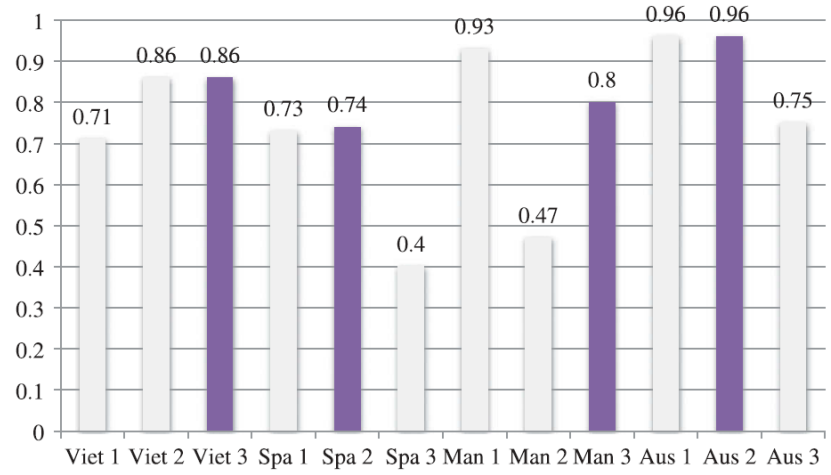


Figure 3. Mean scores of speakers' accent identifiability; the bigger the mean score is the more identifiable the accent is.

The final four speakers underwent recording training to ensure their reading of the test scripts was of similar pitch, speed, and loudness, which was also checked and modified to be consistent by sound-editing software in the final test recordings. They read the listening input texts in a soundproof room with a top-quality recording system. The duration of the four recordings was also controlled to be similar (5 min, ±10 s).

Finalized listening test

The main research instrument for this study was a 30-item Accented English Listening Test (see Appendix C). The texts for the test were chosen from commercial year 8 English test preparation materials sold in China, which was in line with test-takers' background and language proficiency. Researchers and cooperating teachers checked the texts to ensure that they did not contain words or grammatical structures unfamiliar to the target sample. The three task types contained in the test, as displayed in Table 2, are task types frequently encountered by test takers in classroom listening exercises.

The overall listening construct focused on propositional and lexical meaning, including lower and higher levels of psycholinguistic processing (Field, 2013). Section 1 tested comprehension of the propositional meaning of isolated sentences, corresponding to a listening activity of higher complexity (Field, 2013) though the explicitness of the input and its context independence simplified the listening task. Section 2 required the identification of information embedded in a longer monologue and also relied on comprehension of propositions, whereas Section 3 asked test takers to identify and reproduce individual words from the input, which represents a low level of processing complexity in Field's (2013) model.

Section	Task Instructions	Structure	Content		Item Numbers
1: Picture recognition	Listen and choose the right picture	Five sentences corresponding to five pictures	Declarative statements such as "It's important to plant trees."	Sentence length: 10–15 words	1–5, 16–20
2: True/false judgment	Listen to the passage and tell whether the following statements are true or false	One monologue for items 6–10 and another monologue for items 21–25	One on a domestic quibble and the other on giving up smoking	Monologue Length: 70–80 words	6–10, 21–25
3: Gap completion	Listen to the passage and complete the following sentences	One monologue for items 11–15 and another monologue for items 26–30	One on the importance of the sea and the other on a student named Peter and his friends	Monologue Length: 180–200 words	11–15, 26–30

Table 2. Structure of the test.

We also ensured that construct-irrelevant variance possibly introduced through the tasks was limited. While Section 2 involved reading of true/false statements and Section 3 involved productive abilities, the reading items in Section 2 used language that was not likely to pose any comprehensibility challenge for students and the fill-in words in Section 3 were words that students were frequently exposed to.

Questionnaire

To elicit test-takers' subjective perceptions of the accented texts, an Accent Perception Questionnaire was designed. (See Appendix D for the English translation.) The three subscales in the questionnaire were accent familiarity (items 1, 4, 7, 10, 13), accent comprehensibility (items 2, 5, 8, 11, 14), and accent attitude (items 3, 6, 9, 12, 15). The questionnaire was composed in Chinese, the test-takers' first language, to ensure unproblematic comprehension. Items in the questionnaire were statements to which test takers were asked to indicate their degree of agreement across a 7-point scale ranging from "strongly agree" to "strongly disagree." Responses were coded from 1 to 7, with greater familiarity, stronger comprehensibility, and a more positive attitude receiving higher scores. Four negatively phrased items (items 5, 8, 11, 15) were reverse coded. The item results for each subscale were totaled and treated as interval data.

Procedure

The test and questionnaire were administered to the 80 test takers after school hours in the participating middle school in China. All four groups received the same test tasks and questionnaire; the only difference was that each group received listening input with a different accent, similar to the design of Ockey and French's (2014) study.

The four groups of test takers were placed in four different classrooms, each of which was supervised by two English teachers. To ensure reliable scores, test papers were marked by two different raters. In the marking rubric for the test, Sections 1 and 2 consisted of objectively scored items, which allowed for only one correct answer for each question. Section 3 involved filling in missing words, so any semantically equivalent answer was given a full score. While spelling errors were not penalized, grammatical mistakes were penalized with no points given if they concerned incorrect renditions of input. For example, in item 14 ("Today more and more people are _____ the sea.") the grammatically incorrect answer "study" was penalized because it indicates a lack of accurate perception and processing of a grammatical morpheme in the listening input rather than a production error. However, this only occurred on a very small number of occasions.

Finally, the results from the test and questionnaire were coded and entered into SPSS spreadsheets by researchers and double-checked by another volunteer.

Data analysis

Prior to the main data analysis, we performed a psychometric analysis of the test because item characteristics were unknown. We eliminated items with zero variance or insufficient discrimination, retaining a total of 21 items out of the original pool of 30 items. All items were scored dichotomously with items and sections weighted equally in computing section and total scores. Section instructions, items, and section characteristics are shown in Table 3.

The mean score for the final set of 21 items was 58.9%, and Cronbach's alpha reliability was $\alpha = 0.69$, which was considered reasonable for a test with no stakes and a highly homogenous test- taker sample. While the reliabilities of individual sections fluctuated, this is likely due to the small number of items per section and the homogeneity of the test-taker population.

We also investigated the questionnaire items by running correlations among items and Cronbach's alpha reliability for the whole questionnaire and our three subscales Familiarity, Comprehensibility, and

	Items Retained for Final Analysis	Mean	Reliability
Section 1: Picture recognition	6 items: 1, 2, 3, 16, 18, 20	68.5%	α = .45
Section 2: True/False judgment	6 items: 6, 9, 22, 23, 24, 25	78.3%	α = .30
Section 3: Gap completion	9 items: 11, 12, 13, 14, 15, 27, 28, 29, 30	39.6%	α = .55
Whole test	21 items	58.9%	α = .69

Table 3. Sections and their characteristics.

Attitude. We subsequently deleted items showing low correlations with other items on the scale and retained the following items as shown in Table 4.

Subscale inter-item correlations were comfortably within the range of .15 to .5 suggested by Clark and Watson (1995). The reliability of the revised Questionnaire consisting of 10 items was α = 0.82. The overall and subscale reliabilities imply that there is a strong relationship between all items, which would be expected because they are all intended to measure accent perception.

Results

Research question 1a: Overall accent effect

To investigate accent effects, a one-way ANOVA was run on the test scores with the accent group as the factor and the test scores as the dependent variable. Table 5 illustrates descriptive statistics, which show that the Mandarin-accented group scored highest, followed by the Spanish, Australian, and Vietnamese- accented groups. ANOVA results indicate a significant difference between groups, with the factor Group accounting for an appreciable amount of variance ($F_{(3,76)} = 5.796$, $p = .001$, $\eta 2 = 0.19$).

Because of unequal error variances, a Tamhane post hoc test was conducted, and between-group significance levels and effect sizes are reported in Table 6. The group exposed to the Mandarin accent had significantly higher scores than the groups who listened to the Vietnamese and Australian-accented input with large effect sizes for the Mandarin–Vietnamese ($d = 1.245$) and the Mandarin–Australian ($d = 1.098$) comparisons. Although the difference between the Mandarin and the Spanish-accented input was not significant, the significance level for the Mandarin and Spanish comparison ($p = .404$) is smaller

Subscale	Items Retained	Reliability	Inter-item Correlation
Familiarity	items 1, 4, 7	α = .68	.412
Comprehensibility	items 2, 5, 11, 14	α = .76	.439
Attitude	items 3, 9, 12	α = .72	.444

Table 4. Questionnaire subscales.

Group	*N*	*M*	*SD*
Mandarin	20	69.29	14.85
Spanish	20	60.0	17.91
Australian	20	55.48	9.80
Vietnamese	20	50.95	14.58

Table 5. Descriptive statistics for the test.

Accent Groups		*p*	*D*
Mandarin	Vietnamese	.002	1.245
	Australian	.009	1.098
	Spanish	.404	0.564
Vietnamese	Australian	.833	0.179
	Spanish	.425	0.554
Australian	Spanish	.909	0.155

Table 6. Post hoc test on the test results.

than the levels among the three non-Mandarin comparisons (p = .833, .425, .909), and the effect size is medium (d= 0.564), larger than the effect sizes for any of the comparisons not involving Mandarin accent.

Research question 1b: Accent effect and task

Having found an accent effect, we turn to RQ1b to investigate whether this effect differentially impacts different tasks. We ran separate ANOVAs for the three sections of our test: recognize pictures based on individual sentences, judge statements as true or false based on a monologue, and fill in words based on a monologue. Table 7 shows the results for the picture recognition task, and a significant result for the factor Group ($F_{(3,76)}$ = 3.193, p= .028, η2 = 0.12) is observed. While a Scheffé post hoc test did not yield significant differences between groups, the descriptive results suggest that the Mandarin and Spanish accent groups outperformed the Australian and Vietnamese accent groups.

Table 8 shows the results for true/false statements and again, the Mandarin group scored highest, followed by the Spanish group, with the Australian and Vietnamese groups attaining lower scores. ANOVA narrowly missed significance for Group ($F_{(3,76)}$ = 2.693, p = .052, η2 = 0.09).

The results for gap completion tasks are shown in Table 9. As in the previous analyses, the Mandarin group scored highest, followed by the Australian, Spanish, and Vietnamese groups. ANOVA was significant for Group ($F_{(3,76)}$ = 4.657, p= .005, η2 = 0.16) with the strongest effect size of the three task types. A Tamhane post hoc test (run because of unequal error variances) showed that the Mandarin group scored significantly higher than the Vietnamese and Australian groups but did not differ significantly from the Spanish group. None of the non-Mandarin groups differed significantly from each other.

Overall, these results indicate a differential impact of task type. Where test takers had to accurately perceive specific words, the shared-L1 effect was stronger than with tasks that relied more on understanding whole propositions.

Research question 2: Accent and perceptions

To investigate the effect of different accents on test-takers' subjective perception, one-way ANOVAs were conducted separately on Familiarity, Comprehensibility, and Attitude.

Group	*N*	*M*	*SD*
Mandarin	20	76.66	22.55
Spanish	20	75.83	19.09
Australian	20	61.66	20.30
Vietnamese	20	60.00	26.71

Table 7. Section 1 (picture recognition) scores by group.

Group	*N*	*M*	*SD*
Mandarin	20	86.66	13.89
Spanish	20	79.16	20.14
Australian	20	76.66	15.67
Vietnamese	20	70.83	20.85

Table 8. Section 2 (true/false) scores by group.

Group	*N*	*M*	*SD*
Mandarin	20	52.77	20.98
Australian	20	37.22	13.13
Spanish	20	36.66	23.11
Vietnamese	20	31.66	17.01

Table 9. Section 3 (gap completion) by group.

Familiarity

Descriptive statistics for Familiarity are presented in Table 10; they show that test takers judged the Spanish accent as most familiar, with the Mandarin accent judged as nearly equally familiar, followed by Australian and Vietnamese. While ANOVA detected a significant effect for the factor Group ($F_{(3,76)} = 2.846$, $p = .043$, $\eta 2 = 0.1$), a post hoc Scheffé test did not find significant differences between any of the groups. The only difference approaching significance ($p = .073$) was between the Spanish and Vietnamese groups.

While the higher familiarity rating for the Spanish accent may be surprising, it is worth noting that the difference between Spanish and Mandarin accent is clearly nonsignificant ($p = .968$). Test takers also expressed overall a neutral stance toward familiarity with an average rating of 4 (neutral) being the single most common average rating (for 36.3% of test takers) and ratings between 3 and 5 occurring for a clear majority (70.1%).

Comprehensibility

Table 11 illustrates the descriptive statistics for Comprehensibility, which shows that test takers found the Mandarin accent most comprehensible, followed by Spanish, Vietnamese, and Australian. It is interesting that just like the test results, the mean score of the Mandarin group in Comprehensibility is clearly higher than the second highest group's score. Test takers tended toward slight agreement regarding the comprehensibility of Mandarin-accented input, were neutral toward the comprehensibility of Spanish-accented input, and slightly disagreed that input with an Australian or Vietnamese accent was comprehensible.

ANOVA indicates that the between-group difference is significant ($F_{(3,73)} = 6.921$, $p < .001$, $\eta 2 = 0.21$). A Scheffé post hoc test was conducted and the results are reported in Table 12. There is a significant difference in the Mandarin–Vietnamese pair ($p = .005$, d = 1.27) and the Mandarin–Australian pair

Group	*N*	*M* (out of 7)	*SD*
Spanish	20	4.22	1.01
Mandarin	20	4.02	1.41
Australian	20	3.6	1.62
Vietnamese	20	3.15	1.07

Table 10. Descriptive statistics for familiarity (on a scale of 1–7 with higher values indicating greater perceived familiarity).

Group	*N*	*M* (out of 7)	*SD*
Mandarin	20	4.46	1.01
Spanish	19	3.83	1.00
Vietnamese	18	3.24	0.91
Australian	20	3.21	1.13

Table 11. Descriptive statistics for comprehensibility (on a scale of 1–7 with higher values indicating greater perceived comprehensibility).[5]

Accent Groups		*p*	*D*
Mandarin	Vietnamese	.005	1.27
	Australian	.003	1.17
	Spanish	.634	0.63
Vietnamese	Australian	1.0	0.03
	Spanish	.367	0.62
Australian	Spanish	.32	0.58

Table 12. Post hoc test on comprehensibility.

[5] Two test takers in the Vietnamese group and one in the Spanish group did not provide answers to all questionnaire items relating to comprehensibility and were excluded from calculations.

(p= .003, d= 1.17), and the effect sizes for the two pairs are large. No significant difference is found in the Mandarin–Spanish pair or any of the non-Mandarin pairs, again echoing the test findings.

Attitude

Descriptive statistics for Attitude are presented in Table 13. Test takers from the Spanish-accented input group responded most positively, followed by Mandarin, Vietnamese, and Australian English. The difference between the most favored and the least favored accent is small, and ANOVA analysis further confirms that it is nonsignificant ($F_{(3,76)} = 0.581$, $p = .63$). Attitude judgments were the only category where all test-taker groups clustered around slight agreement, indicating an overall positive attitude to accented input.

Research question 3: Correlation of questionnaire subscales with test scores

Pearson correlation was run on the three questionnaire subscales with the test scores, findings of which are presented in Table 14. Comprehensibility (r = .471) and Familiarity (r = .303) correlate moderately but significantly with test scores, whereas the correlation with Attitude (r= .219) is weaker and nonsignificant. The correlation between Comprehensibility and test scores indicates that test-takers' subjective perception of the comprehensibility of the test predicts their performance to some extent, with better self-perceived Comprehensibility related to higher test scores. By comparison, test-takers' perceived familiarity was a weaker predictor, and attitude toward accents did not predict their test performance well.

Discussion

Accent and test performance

Similar to Harding (2011) and Major et al. (2002), this study also found partial support for a shared-L1 advantage phenomenon. Our study reflects the findings of both Harding (2011) and Major et al. (2002) where a shared accent effect was observed for some groups but not others. Similar to Harding's (2011) study but in contrast to Major et al.'s (2002) results, test takers listening to Mandarin-accented input showed a shared-L1 effect or interlanguage speech intelligibility benefit (Bent & Bradlow, 2003), out- performing test takers receiving non-L1 accented input. This difference was not significant for Spanish, which could be interpreted as evidence of Bent and Bradlow's (2003) mismatched interlanguage speech intelligibility benefit, but the picture here is less clear. The three non-Mandarin groups' scores were not significantly different for the total test, which indicates that the non-Mandarin-accented tests were of similar difficulty, whether they were recorded in a native English accent (Australian), a nontonal non- native accent (Spanish), or a tonal non-native accent (Vietnamese). However, Spanish-accented input seems to have had less of a detrimental effect on comprehension, pointing to factors other than tonality and nativeness as explanations for accent effects in our study. Major et al. (2002) also found a beneficial

Group	*N*	*M* (out of 7)	*SD*
Spanish	20	5.38	1.04
Mandarin	20	5.03	1.60
Vietnamese	20	5.02	1.10
Australian	20	4.92	.92

Table 13. Descriptive statistics for attitude (on a scale of 1–7 with higher values indicating more positive attitude).

Correlation		*N*	*r*	*p*
Test scores	Familiarity	80	.303	.006
	Comprehensibility	78	.471	<.001
	Attitude	80	.219	.051

Table 14. Correlation of questionnaire subscales with test scores.

effect of Spanish-accented input for Chinese listeners, and test-takers' high-familiarity ratings for the Spanish accent indicate that phonological features of Spanish seem to aid comprehension. Conversely, there might be inherent difficulties in comprehending Vietnamese-accented English due to the deletion of final English consonants and reduction of English consonant clusters (Derwing, Rossiter, & Munro, 2002; Hultzén, 1965). These differences between L2 accents are not likely to have serious consequences in real-world testing settings, where several non-native accents might occur in the test so that their effects would most likely even out.

An interesting point that might benefit from further investigation is that the shared-L1 effect seems to impact various task types differently. The facilitative effect of the shared accent was strongest for gap completion items where test takers needed to isolate single words from the input stream. This may be due to segmentation being more familiar and phonemes appearing less ambiguous with a shared accent, which is particularly important for the comprehension of individual words because filling gaps becomes guesswork if the target lexical item is not fully understood in the input. By contrast, where the comprehension of larger propositions is required, the accurate comprehension of individual items is less important. However, the interaction between accent and task deserves further exploration.

Comprehensibility, familiarity, attitude, and test performance

Turning to test-taker impressions, it is not surprising that test takers reported that they found the Mandarin-accented input most comprehensible. This further validates the finding of a shared-accent effect from a test-taker perspective, as does the midstrength and significant correlation between comprehensibility judgments and test scores. Test-takers' judgment of the comprehensibility of the three non-Mandarin accents also corresponds with their performance in the test with Mandarin ranked first, Spanish second, followed by Vietnamese and Australian. This congruence between scores and comprehensibility judgments is probably due to test takers being able to base judgments on their recent test-taking experience. It is also interesting that the native Australian accent neither facilitated test-takers' performance in the test, nor was it considered more comprehensible from test-takers' subjective perception in the comprehensibility scale. It therefore appears that native-accented listening input does not lead to a different test-taker experience compared to non-native-accented input, although it is possible that using a dominant variety of English might have a stronger effect.

For test-taker impressions of familiarity, results from this study suggest that although a test taker may believe that they have some degree of familiarity with an accent, familiarity does not translate strongly into better performance on the test. This seems to contradict findings by Ockey and French (2014), who found a generally facilitative effect of familiarity. In their study, familiarity aided comprehension of British English input regardless of accent heaviness but interacted with accent heaviness for Australian English input. Ockey and French attributed this discrepancy to the fact that self-perceived familiarity with accents could be unreliable because test-takers' familiarity assessments are likely based on vague impressions of their prior exposure to an accent but their actual familiarity may be greater or lesser than their own estimate. Similarly, Ballard and Winke (2017) showed that non-native speakers of English generally find it difficult to determine the origin of an accent in English, which is in accordance with our finding that a strong majority of test takers clustered around a "neutral" judgment of familiarity, indicating that they were unable to come to a clear determination of their familiarity with the four accents. This uncertainty in familiarity judgments likely accounts for the small effect of familiarity in our study as well as test-takers' judgment of the Spanish accent as the most familiar one, though not significantly more familiar.

However, even though they may not be aware of it, test takers were likely familiar with the segmental and suprasegmental features in the accent of a speaker with a shared L1, which improved their comprehension of a listening test recorded in this speaker's accent. In this study, although subjectively test takers failed to recognize Mandarin accent as the most familiar one and their familiarity judgments only have low correlation with test scores, they still benefited from the shared- L1 effect, which translated into better test performance.

Similar to familiarity, there was also no significant difference between accent groups in attitude judgments. The nonsignificant between-group difference on the attitude scale shows that test takers in this study did not have any particular preference for any of the four accents at a group level. This result challenges previous findings that listeners prefer the native accents of a certain language to non-native accents (Fraser & Kelly, 2012; Hendriks et al., 2015; Hendriks, van Meurs, & van der Meij, 2015), and it may be due to a combination of factors. In this study, test takers did not compare accents against each other, as they did in Butler's (2007) study, where participants listened to both types of accented input when answering the attitude questionnaire. This likely made it easier to distinguish the native-accented input from the non-native accented input, whereas our participants only judged one type of accented input. The findings from the familiarity questionnaire indicate that they probably did not guess speakers' backgrounds, given that they considered the Spanish accent slightly more familiar than the Mandarin accent. When unaware of speakers' first language background, test takers would be less concerned with accents' social status (Nejjari et al., 2012). Other issues, such as distorted accent perception (Hu & Lindemann, 2009) and perception bias resulting from accent identification (Atagi & Bent, 2016; McKenzie, 2015; Winke & Gass, 2013), are also circumvented. Therefore, it appears that inexperienced listeners perceive various accents as equally favorable when they are not influenced by extraneous factors. It also shows that preferences for certain accents are largely the result of social learning: Our test takers were not informed of speakers' nationality and showed no partiality toward any of the accents, whether it was their L1 accent, a native accent, or other non-native accents.

Limitations and implications

Our study had some limitations. We did not include a British accent even though students had frequent exposure to it because the teaching materials were based on British English. While using a British accent could have affected their accent familiarity for the native English variety, the focus of our study was on non-native accents, so our central findings of a shared-L1 effect and test-takers' positive attitude toward non-native accents would have been unaffected. Another limitation was the modest number of items for each task type in the test and in the questionnaire. More items would likely have contributed to higher reliability and better understanding of the relationship between accent and task as well as accent and test-taker perceptions.

For implications, findings from this study have offered some support for the shared-L1 advantage argument, and it is clear that this issue will need to be addressed for the inclusion of non-native English accents into high-stakes international English listening assessments. Possible solutions for tackling the shared-L1 issue have been suggested in Harding (2011), such as only including the most frequently used non-native accents in a given context, adopting highly intelligible non-native accents, and balancing non-native accents across various listening tasks. Our findings on the differential effect of shared accent for different tasks suggest that the shared-L1 effect is possibly less of a concern with tasks that require comprehension of larger propositions than with tasks that focus on individual words, so the inclusion of various task types can also help dilute the shared-L1 effect. In addition, while it needs to be controlled for reasons of test fairness, a shared-L1 advantage is not entirely construct-irrelevant because test takers would likely enjoy the same advantage in real- world interaction.

Given our finding that there is no statistical difference in either test-takers' performance or their evaluation of comprehensibility in the three non-L1 accents, this study provides support for multidialectal listening assessments. The implication is that an international listening test can include a variety of accents, and they do not seem to adversely affect comprehension, at least where these accents are mild. Future research will need to determine if this lack of accent effect holds across test-taker native languages or if certain accents are inherently more difficult for test takers from specific L1 backgrounds. It also would be important to investigate whether speakers' degree of accent strength interacts with speaker native language and test-taker native language in

performance and subjective impressions. Furthermore, effects of age should be systematically investigated because our test takers were middle school adolescents, and primary school learners may react differently to accents, which might also account for the differences between our findings and Butler's (2007).

Finally, the absence of any significant differences in test-taker attitudes toward non-L1 accents and the weak effect of attitude on performance is an encouraging finding for test developers, who are understandably concerned with the acceptability and uptake of a new test feature in the marketplace. The test format and item types in this study embody more traditional listening constructs instead of a more radical EFL turn as suggested in Elder and Davies (2006), and this conservative approach can smooth the transition to a new testing paradigm, which tends to be a source of concern for stakeholders with vested interests (Harding & McNamara, 2017).

Conclusion

This study is situated in the ELF debate and examined how adolescent test takers performed in a listening test when non-native accents were introduced. Our findings point toward a shared-L1 effect on both the test scores and test-takers' subjective impression of comprehensibility, although this effect is not clear-cut with scores on one of the non-L1 accents not significantly different from the L1 accent. At the same time, there is no significant difference in test-takers' performance in the three non-L1 accents or in their self-reported familiarity and attitude. This finding supports the inclusion of multidialectal listening assessment tasks to better reflect the actual multidialectal status of English as an international language.

Acknowledgments

The authors thank the editor and the anonymous reviewers of Language Assessment Quarterly for their detailed and insightful comments on an earlier draft of this article. The authors also thank the English teachers and students in China who participated in this study.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

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

Adank, P., Evans, B. G., Stuart-Smith, J., & Scott, S. K. (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. Journal of Experimental Psychology: Human Perception and Performance, 35(2), 520–529. doi:10.1037/a0013552

Atagi, E., & Bent, T. (2016). Auditory free classification of native and nonnative speech by nonnative listeners. Applied Psycholinguistics, 37(2), 241–263. doi:10.1017/S014271641400054X

Ballard, L., & Winke, P. (2017). Students' attitudes towards English teachers' accents: The interplay of accent familiarity, comprehensibility, intelligibility, perceived native speaker status, and acceptability as a teacher. In

T. Isaacs & P. Trofimovich (Eds.), Second language pronunciation assessment: Interdisciplinary perspectives (pp. 121–140). Bristol, UK: Multilingual Matters.

Bent, T., & Bradlow, A. R. (2003). The interlanguage speech intelligibility benefit. The Journal of the Acoustical Society of America, 114(3), 1600–1610.

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

British Council. (2016). English test for teens. Retrieved from https://www.britishcouncil.org/exam/aptis/assessment/ school-university/aptis-teens

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

Butler, Y. G. (2007). How are nonnative-English-speaking teachers perceived by young learners? TESOL Quarterly, 41(4), 731–755. doi:10.1002/tesq.2007.41.issue-4

Chien, S. (2014). Varieties of English: Taiwanese attitudes and perceptions. Newcastle and Northumbria Working Papers in Linguistics, 20, 1–16.

Cho, Y., Ginsburgh, M., Morgan, R., Moulder, B., Xi, X., & Hauck, M. C. (2016). Designing the TOEFL® primary™ tests. (Research Memorandum No. RM-16-02). Princeton, NJ: Educational Testing Service.

Clark, L. A, & Watson, D. B. (1995). Constructing validity: Basic issues in objective scale development. Psychological Assessment, 7, 309–319.

Department of Education and Training, Australian Government. (2014). International student data 2014. Retrieved fro m https://internationa l educat ion.gov.a u/researc h /International -Student-Data/Pa ges/ InternationalStudentData2014.aspx

Derwing, T. M., & Munro, M. J. (2015). Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research. Amsterdam, the Netherlands: John Benjamins.

Derwing, T. M., Rossiter, M. J., & Munro, M. J. (2002). Teaching native speakers to listen to foreign-accented speech.

Journal of Multilingual and Multicultural Development, 23(4), 245–259. doi:10.1080/01434630208666468

Elder, C., & Davies, A. (2006). Assessing English as a lingua franca. Annual Review of Applied Linguistics, 26, 282–304. doi:10.1017/S0267190506000146

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, UK: Cambridge University Press.

Fraser, C., & Kelly, B. (2012). Listening between the lines: Social assumptions around foreign accents. Australian Review of Applied Linguistics, 35(1), 74–93. doi:10.1075/aral

Gass, S., & Varonis, E. M. (1984). The effect of familiarity on nonnative speech. Language Learning, 34, 65–89.
Harding, L. (2011). Accent, listening assessment and the potential for a shared-L1 advantage: A DIF perspective.

Language Testing, 29(2), 163–180. doi:10.1177/0265532211421161

Harding, L., & McNamara, T. (2017). Language assessment: The challenge of ELF. In J. Jenkins, W. Baker, &

M. Dewey (Eds.), The routledge handbook of English as a lingua franca (pp. 570–582). New York, USA: Routledge.
Hendriks, B., Meurs, F., & Groot, E. (2015). The effects of degrees of Dutch accentedness in ELF and in French,

German and Spanish. International Journal of Applied Linguistics. doi:10.1111/ijal.12101

Hendriks, B., van Meurs, F., & van der Meij, E. (2015). Does a foreign accent sell? The effect of foreign accents in radio commercials for congruent and non-congruent products. Multilingua, 34(1), 119–130.

Hiraga, Y. (2005). British attitudes towards six varieties of English in the USA and Britain. World Englishes, 24(3), 289–308. doi:10.1111/weng.2005.24.issue-3

Hu, G., & Lindemann, S. (2009). Stereotypes of Cantonese English, apparent native/non-native status, and their effect on non-native English speakers' perception. Journal of Multilingual and Multicultural Development, 30(3), 253–269. doi:10.1080/01434630802651677

Hultzén, L. S. (1965). Consonant clusters in English. American Speech, 40(1), 5–19. doi:10.2307/454173

Institute for International Education. (2014). Charting new pathways to higher education: International secondary students in the United States. Retrieved from http://www.iie.org/~/media/Files/Corporate/Publications/IIE-International-Secondary-Students-In-The-US.pdf.

Jenkins, J., & Leung, C. (2013). English as a Lingua Franca. In A. J. Kunnan (Ed.), The companion to language assessment (pp. 1605–1616). West Sussex, UK: John Wiley & Sons.

Kim, Y. S. (2007). Korean adults' attitudes towards varieties of English. Unpublished manuscript, University of Edinburgh, Edinburgh, United Kingdom.

Lindemann, S. (2017). Variation or 'error'? Perception of pronunciation variation and implications for assessment. In

T. Isaacs & P. Trofimovich (Eds.), Second language pronunciation assessment: Interdisciplinary perspectives (pp. 193–209). Bristol, UK: Multilingual Matters.

Lobo & Yoshida. (1982). The perceptual acquisition of English phonology by Japanese students. Paper presented at Georgetown University Roundtable on Languages and Linguistics, Washington, DC.

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

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

McKenzie, R. M. (2008). Social factors and non-native attitudes towards varieties of spoken English: A Japanese case study. International Journal of Applied Linguistics, 18(1), 63–88. doi:10.1111/ijal.2008.18.issue-1

McKenzie, R. M. (2015). The sociolinguistics of variety identification and categorisation: Free classification of varieties of spoken English amongst non-linguist listeners. Language Awareness, 24(2), 150–168. doi:10.1080/ 09658416.2014.998232

Munro, M. J, & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. Language Learning, 45(1), 73–97.

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. doi:10.1017/S0272263106060049

Nejjari, W., Gerritsen, M., Van der Haagen, M., & Korzilius, H. (2012). Responses to Dutch-accented English. World Englishes, 31(2), 248–267. doi:10.1111/weng.2012.31.issue-2

Ockey, G. J., & French, R. (2014). From one to multiple accents on a test of L2 listening comprehension. Applied Linguistics. Advanced online publication. doi:10.1093/applin/amu060

Pica, T., & Long, M. (1982, May). The classroom linguistic and conversational performance of experienced and inexperienced teachers. Paper presented at 16th annual TESOL convention, Honolulu, HI.

Seargeant, P. (2012). Disciplinarity and the study of world Englishes. World Englishes, 31(1), 113–129. doi:10.1111/ weng.2012.31.issue-1

Seidlhofer, B. (2011). Understanding English as a lingua franca. Oxford, UK: Oxford University Press.

Sidaras, S. K., Alexander, J. E., & Nygaard, L. C. (2009). Perceptual learning of systematic variation in Spanish-accented speech. The Journal of the Acoustical Society of America, 125(5), 3306–3316. doi:10.1121/ 1.3101452

So, Y., Wolf, M. K., Hauck, M. C., Mollaun, P., Rybinski, P., Tumposky, D., & Wang, L. (2015). TOEFL junior® design framework. ETS Research Report Series, 2015(1), 1–45. doi:10.1002/ets2.2015.2015.issue-1

Stevenage, S. V., Clarke, G., & McNeill, A. (2012). The “other-accent” effect in voice recognition. Journal of Cognitive Psychology, 24(6), 647–653. doi:10.1080/20445911.2012.675321

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.

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

University of Cambridge Local Examinations Syndicate (2016). Cambridge English: Young learners. Retrieved from http://www.cambridgeenglish.org/exams/young-learners-english/

Varonis, E., & Gass, S. (1982). The comprehensibility of non-native speech. Studies in Second Language Acquisition, 4(2), 114–136. doi:10.1017/S027226310000437X

Weber, A., Broersma, M., & Aoyagi, M. (2011). Spoken-word recognition in foreign-accented speech by L2 listeners.Journal of Phonetics, 39(4), 479–491. doi:10.1016/j.wocn.2010.12.004 Weinberger, S. (2017). Speech accent archive. Retrieved from http://accent.gmu.edu

Winke, P., & Gass, S. (2013). The influence of second language experience and accent familiarity on oral proficiency rating: A qualitative investigation. TESOL Quarterly, 47(4), 762–789. doi:10.1002/tesq.73

Yook, C., & Lindemann, S. (2013). The role of speaker identification in Korean university students‘ attitudes towards five varieties of English. Journal of Multilingual and Multicultural Development, 34(3), 279–296. doi:10.1080/ 01434632.2012.734509

Zhang, Q. (2009). Hong Kong people‘s attitudes towards varieties of English. Newcastle Working Papers in Linguistics, 15, 151–173.

Appendix A: Strength of Accent Scale

Band 1
The speaker’s accent is **very similar to** what I am used to.
I can concentrate on listening **without any problem**.
I can **easily** understand the recording.

Band 2
The speaker’s accent is **slightly different** from what I am used to.
I can concentrate on listening **without too much trouble**.
I can understand the recording **to a large extent**.

Band 3
The speaker’s accent is **different** from what I am used to.
I find it **slightly challenging** to concentrate on listening.
I can **roughly** understand the recording.

Band4
The speaker’s accent is **very different** from what I am used to.
I need to concentrate on listening **more than usual**.
I have **limited** understanding of the recording.

Band 5
The speaker’s accent is **noticeably different** from what I am used to.
I have to **excessively** concentrate on listening.
I can **barely** understand the recording.

Appendix B: Accent Strength and Identification Task

I. The Background Section

1. How would you describe your English listening skills?

☐ very limited ☐ limited ☐ average ☐ good ☐ excellent

Do you think you are good at telling English accents apart?

☐ not at all ☐ limited ☐ average ☐ good ☐ excellent

2. Is English your mother tongue? If not, please write down your mother tongue(s).

(It is possible that you might have more than one mother tongue if you grew up in a multilingual community)

Yes ☐

No ☐

Other mother tongue(s):

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

Australian English accent

☐ no knowledge ☐ a little familiar ☐ average ☐ familiar ☐ very familiar

Spanish English accent

☐ no knowledge ☐ a little familiar ☐ average ☐ familiar ☐ very familiar

Vietnamese English accent

☐ no knowledge ☐ a little familiar ☐ average ☐ familiar ☐ very familiar

Mandarin English accent

☐ no knowledge ☐ a little familiar ☐ average ☐ familiar ☐ very familiar

II. The accent judgment section

Listen to each recording clip, decide the accentedness and identify the accent of each clip.

Clip 1:

☐ Band 1 ☐ Band 2 ☐ Band 3 ☐ Band 4 ☐ Band 5

☐ Australian ☐ Spanish ☐ Vietnamese ☐ Mandarin

Clip 2:

☐ Band 1 ☐ Band 2 ☐ Band 3 ☐ Band 4 ☐ Band 5

☐ Australian ☐ Spanish ☐ Vietnamese ☐ Mandarin

Clip 3:

☐ Band 1 ☐ Band 2 ☐ Band 3 ☐ Band 4 ☐ Band 5

☐ Australian ☐ Spanish ☐ Vietnamese ☐ Mandarin

Clip 4:

☐ Band 1 ☐ Band 2 ☐ Band 3 ☐ Band 4 ☐ Band 5

☐ Australian ☐ Spanish ☐ Vietnamese ☐ Mandarin

Clip 5:

☐ Band 1 ☐ Band 2 ☐ Band 3 ☐ Band 4 ☐ Band 5

☐ Australian ☐ Spanish ☐ Vietnamese ☐ Mandarin

Clip 6:

☐ Band 1 ☐ Band 2 ☐ Band 3 ☐ Band 4 ☐ Band 5

☐ Australian ☐ Spanish ☐ Vietnamese ☐ Mandarin

Clip 7:

☐ Band 1 ☐ Band 2 ☐ Band 3 ☐ Band 4 ☐ Band 5

☐ Australian ☐ Spanish ☐ Vietnamese ☐ Mandarin

Clip 8:

☐ Band 1 ☐ Band 2 ☐ Band 3 ☐ Band 4 ☐ Band 5

☐ Australian ☐ Spanish ☐ Vietnamese ☐ Mandarin

Clip 9:

☐ Band 1 ☐ Band 2 ☐ Band 3 ☐ Band 4 ☐ Band 5

☐ Australian ☐ Spanish ☐ Vietnamese ☐ Mandarin

Clip 10:

☐Band 1 ☐Band 2 ☐Band 3 ☐Band 4 ☐Band 5

☐Australian ☐Spanish ☐Vietnamese ☐Mandarin

Clip 11:

☐Band 1 ☐Band 2 ☐Band 3 ☐Band 4 ☐Band 5

☐Australian ☐Spanish ☐Vietnamese ☐Mandarin

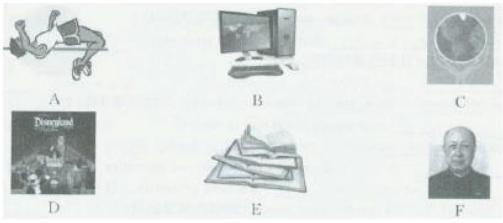
Clip 12:

☐Band 1 ☐Band 2 ☐Band 3 ☐Band 4 ☐Band 5

☐Australian ☐Spanish ☐Vietnamese ☐Mandarin

Appendix C: Accented English Listening Test

I. Listen and choose the right picture.



1. _ _ _ 2 . _ _ _ 3 . _ _ _ 4 . _ _ _ 5 . _ _ _

II. Listen to the passage and tell whether the following statements are true or false.

() 6. The wife was a large woman and the husband was a small man.
() 7. The man worked as a manager in a big company.
() 8. The wife gave her weekly money to her husband.
() 9. The wife would take all the money from her husband and never give him any money.
() 10. The husband was happy to tell his wife that he had won a lot of money.

III. Listen to the passage and complete the following sentences.

11. The sea is the _____ of millions of living things.
12. There is more life in the sea than on _____ .
13. The animals and plants of the sea are very _____ to man as a source of food.
14. Today more and more people are _____ the sea.
15. The land will not be _____ to provide food for everybody.

IV. Listen and choose the right picture.

16. _ _ _ 17 . _ _ _ 18 . _ _ _ 19 . _ _ _ 20 . _ _ _

V. Listen to the passage and tell whether the following statements are true or false.

() 21. Young people smoke because they think it is cool.
() 22. The famous star Jackie Chan also thinks it is cool to smoke.
() 23. Yao Ming isn't so cool because he doesn't smoke cigarettes.
() 24. Some young people smoke because they see their parents do that.
() 25. Exercise is a good way to help us give up smoking.

VI. Listen to the passage and complete the following sentences.

26. Jack has _____ hair and blue eyes.
27. Peter and Jack are both on the school _____ team.
28. Linda won the women's first _____ of their school.
29. Betty came to Peter's class _____ months ago.
30. Peter _____ here from Europe.

Appendix D: Accent Perception Questionnaire

Item	Statements	Strongly Agree	Agree	Slightly Agree	Neutral	Slightly Disagree	Disagree	Strongly Disagree
1	I am familiar with the accent in the test							
2	I think I can understand the accent in the test							
3	I hope in future English classes I can hear more of the accent of the test							
4	I think the accent of the test is similar to the accent of my textbook recordings							
5	The accent of the test impeded my understanding of the test content							
6	I find it necessary to develop the ability to understand different English accents							
7	I think the accent of the test is similar to our English teacher's accent							
8	The accent of the test made it hard for me to concentrate							
9	I like the accent of the test							
10	If the accent of the test were the accent of our textbook recordings, I would have performed better in the test							
11	The accent of the test made the test harder than usual							
12	I hope I can understand different English accents							
13	If the accent of the test were the accent of our English teacher, I would have performed better in the test							
14	Due to the accent, the test became easier							
15	I find the accent of the test very weird							