Accent, listening assessment and the potential for a shared-L1 advantage: A DIF perspective

Luke Harding
Lancaster University

Published as:

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

Available online: http://ltj.sagepub.com/content/29/2/163

Abstract

This paper reports on an investigation of the potential for a shared-L1 advantage on an academic English listening test featuring speakers with L2 accents. 212 second-language listeners (including 70 Mandarin Chinese L1 listeners, and 60 Japanese L1 listeners) completed three versions of the University Test of English as a Second Language (UTESL) listening sub-test which featured an Australian English-accented speaker, a Japanese-accented speaker and a Mandarin Chinese-accented speaker. Differential item functioning (DIF) analyses were conducted on data from the tests which featured L2 accented speakers using two methods of DIF detection – the standardization procedure and the Mantel-Haenszel procedure – with candidates matched for ability on the test featuring the Australian English accented speaker. Findings showed that Japanese L1 listeners were advantaged on a small number of items on the test featuring the Japanese accented speaker, but these were balanced by items which favoured non-Japanese L1 listeners. By contrast, Mandarin Chinese L1 listeners were clearly advantaged across several items on the test featuring a Mandarin Chinese L1 speaker. The implications of these findings for

claims of bias are discussed with reference to the role of speaker accent in the listening construct.

Keywords: listening assessment, accent, intelligibility, listening comprehension, pronunciation, differential item functioning

I Introduction

New perspectives on the use of English as an international language (EIL) have presented significant challenges to the field of language testing, with calls for change in assessment practices arising over the past decade (see Canagarajah, 2006; Jenkins, 2006; Lowenberg, 2002). These critiques have generally focused on the traditional centrality of native speaker norms to assessment standards (see Davies, Hamp-Lyons & Kemp, 2003), and within this broader question, the issue of speaker accent in listening assessment has emerged as an area of inquiry (Llurda, 2004). Currently, many large-scale tests of English such as TOEFL® and IELTS utilise a range of accents in their listening assessment input. In the TOEFL iBT, recent innovations to the listening section have included British and Australian accents in addition to North American varieties, though these 'new' accents are only heard in one part of the listening section (mini-lectures), and may not appear in all listening tests (see ETS, 2005). Similarly, candidates sitting IELTS may encounter British, Australian, North American or New Zealand varieties (see Cambridge ESOL, 2008).

However in many of the target-language use domains for which listening assessments are designed (e.g., academic, health communication, business), listeners are likely to encounter not only a wide range of native speaker accents but also a range of second language (L2) varieties (see Canagarajah, 2006). On this basis, a rationale exists for the inclusion of L2 accents in listening tests on the bases of enhanced authenticity, a more accurate representation of the listening construct, and the potential for positive washback (see Harding, 2011). Yet while the inclusion of L2 accents in listening assessment may be viewed as beneficial, it also raises concerns relating to validity and practicality (for discussion see Taylor, 2006; Taylor & Geranpayeh, 2011). These concerns appear to have driven an orthodox approach in listening test development in which accent choice has traditionally been limited to native speaker varieties.

Prominent among these concerns is the potential for test bias if listeners who share a speaker's first language (L1) are advantaged over others when listening to that speaker (Major et al., 2002; Taylor, 2006). The term 'bias' in language testing has been defined as existing when candidates of equal ability, but from different groups, have an unequal chance of getting an item correct, or of attaining the same test score due to a factor which is not relevant to the construct under test (see Angoff, 1993; Zumbo, 2007). With relation to a listening test featuring L2 speakers, the concern for bias relates to the possibility that a shared-L1 advantage would translate to a systematically improved score performance for test-takers who happen to share the L1 of a speaker compared to other listeners of equal ability who do not.

However, to date there has been little empirical research conducted on the potential for a shared-L1 effect within an assessment context, and the findings of those few studies which have investigated the issue (see below) have been inconclusive. In addition, the implications of a shared-L1 advantage for claims of test bias have not been adequately addressed in a conceptual sense. The aim of this study, then, was to investigate the potential for a shared-L1 advantage for listeners in an academic English listening test featuring speakers with L2 accents, and to assess claims of the potential for test bias in light of these findings.

II Research questions and methodological approach

In order to address this aim, the study addresses the following two research questions:

1) Do test-takers who share a speaker's L1 perform better on a listening test featuring that speaker than test-takers of equal ability who do not?
2) To what extent does evidence of a shared-L1 effect constitute test bias?

III Research on shared-L1 effects

The possibility of a shared-L1 intelligibility advantage has been explored in various fields, notably in cross-language speech perception (see Best, 1995; Strange, 1995) and also in research on World Englishes (e.g., Smith and Rafiqzad, 1979). From the perspective of cross-language speech perception, there is a theoretical foundation for a shared-L1 effect based on the principle that L2 accents are primarily characterised by transfer from the L1, and that listeners who share a speaker's L1 will have an intimate familiarity with the phonological patterns of that speaker's L2 accent. In this respect, Bent and Bradlow (2003) characterise shared-L1 speaker-listener pairs as sharing a knowledge base which includes, 'the system of consonants and vowel categories, phonotactics, stress patterns, and intonation as well as other features of the sound system' (p.1607). From a World Englishes perspective, a shared-L1 intelligibility advantage would form evidence of common pronunciation norms in an emerging variety. Robust empirical findings have shown that exposure to an accent aids intelligibility of a different speaker with the same accent (see Bradlow & Bent, 2008; Clarke & Garrett, 2004), and it has been assumed that L2 listeners will find speakers who share their L1 most comprehensible given their greater level of familiarity with that variety (Flowerdew, 1994).

However, although a spate of empirical research has addressed the specific issue of a shared-L1 intelligibility advantage, findings have been mixed. A significant contribution in this area was Bent and Bradlow's (2003) study in which perception tests were conducted with Chinese, Korean, English native speaker and mixed-nationality groups listening to speakers with Chinese, Korean and English native speaker accents reading sentences in English. Results showed that the native listeners found native-speakers of English the most intelligible; however, for each of the non-native listener groups, a highly proficient non-native speaker from the same L1 background was as intelligible as a native-speaker. Bent and Bradlow labelled this phenomenon a 'matched interlanguage intelligibility benefit' (p.1606), however this label is slightly misleading in its suggestion of advantageous conditions for shared-L1 talker-listener pairs when in most cases results showed only *equivalent* levels of intelligibility for native-speakers and highly proficient non-native speakers. Bent and Bradlow, though, also found that a low proficiency Korean speaker was of a similar intelligibility for Korean listeners as the high proficiency Chinese speaker and the native speaker, lending weight to their argument. In a replication study, Stibbard and Lee (2006) also found that within each listener group there were no significant differences in intelligibility scores for high proficiency non-native speakers and native speakers. While in each of these studies a speaker-listener interaction effect was significant, the findings suggest that a shared-L1 effect may only be taking hold when listeners hear lower proficiency speakers.

Taking a different stance, Munro, Derwing and Morton (2006) have demonstrated that a shared-L1 background may have little impact on intelligibility. Munro et al. measured the intelligibility of Japanese, Cantonese, Polish and Spanish L1 speakers for listeners from Japanese, Cantonese, Mandarin and Canadian English language backgrounds. There was a significant interaction effect between speaker and listener

groups, although the effect size for speaker on its own was much larger. Post hoc analyses revealed that Japanese listeners found the Japanese speaker more intelligible than any of the other listener groups. However this effect did not hold when a Cantonese shared-L1 advantage was investigated. Munro et al. stress that this provides only weak evidence of a shared-L1 effect, and hypothesise that the effect is so small that it is 'readily outweighed by other factors' (2006, p.127).

The potential for a shared-L1 advantage has also been investigated in studies of second language listening comprehension (Ortmeyer & Boyle, 1985; Smith & Bisazza, 1982; Tauroza & Luk, 1997; Yule, Wetzel & Kennedy, 1990). Across these studies, while a shared-L1 advantage was observed for particular L1 speaker-listener pairs, the phenomenon was inconsistent. However two general conclusions have emerged from this research. Firstly, it has been hypothesised that some L1 groups find their own accent more comprehensible because the sociolinguistic *milieu* dictates that their own variety is the most familiar, while for other L1 groups, a particular native speaker variety might be equally as familiar as a shared-L1 accent (Smith & Bisazza, 1982; Tauroza & Luk, 1997). This may in part explain the inconsistent results across these studies. The second conclusion is that shared-L1 effects might be less noticeable when a L2 speaker has a high level of general intelligibility (Ortmeyer & Boyle, 1985), reflecting, to some extent, the findings of Bent and Bradlow (2003).

Within the field of language testing, one paper of particular pertinence to this investigation is Major et al.'s (2002) study of the potential for a shared-L1 effect on scores for TOEFL listening tasks. Major et al. conducted an experiment in which TOEFL PBT mini-talk lectures were delivered in English by speakers with Chinese, Japanese, Korean and Spanish L1 backgrounds. Tests were then administered to a cohort of listeners who shared the same L1 backgrounds. A repeated-measures ANOVA showed a significant

interaction effect between speaker and listener L1 background, and post hoc analyses showed that Spanish learners were advantaged by listening to a Spanish L1-background speaker, but the Chinese listener group was disadvantaged by listening to a Chinese-accented speaker. This led Major et al. to conclude that listeners 'sometimes' performed better when listening to a shared-L1 speaker on a listening test (2002, p.185), reflecting the mixed results of preceding studies. However, based on their results, Major et al. recommend a cautious approach, arguing that using L2 varieties in listening test materials 'may create test bias, thereby posing a threat to construct validity' (p.188).

There are two limitations, though, in the methodology of the Major et al. study. Firstly, as the researchers acknowledge in their paper, the results are affected by the possible incomparable difficulty of listening comprehension tasks. Secondly, the design of the study addresses the issue of bias only indirectly. Given the definition of bias articulated at the beginning of this paper which relates to differential group performance, it is necessary to investigate bias by focusing on between-groups differences. Major et al. found an interaction effect between speaker and L1 listener group, but only explain their findings by comparisons of different speaker effects within each listener group. These patterns suggest that listener groups varied in their responses to each speaker, but they do not give clear evidence of test bias.

IV Differential item functioning (DIF) and bias

In order to investigate the research questions, the current study drew on differential item functioning (DIF), a common approach used in the language testing literature to investigate bias. Differential item functioning is generally defined as existing when two groups of test-takers, who are otherwise matched in ability on a construct, have different probabilities of answering an item correctly (see Ferne & Rupp, 2007). A DIF finding,

which in essence signifies the advantage of one group over another, may be attributed to the influence of construct-irrelevant variance on the studied item (and so indicate 'item bias'). On the other hand, two groups may differ in a construct-relevant way, in which case DIF may indicate impact rather than bias. DIF is therefore regarded as 'a necessary but not sufficient condition' for establishing an argument for bias (McNamara & Roever, p.83).

Various procedures have been used to calculate DIF, and according to McNamara and Roever (2006) these can be classified into four categories: analyses based on item difficulty, nonparametric approaches, item response theory (IRT) approaches, and 'other' approaches (such as logistic regression). These approaches have emerged more or less chronologically, with item difficulty approaches often found in early DIF studies, and IRT and logistic regression appearing more recently. Each 'family' of approaches has different strengths and assumptions. Ferne and Rupp (2007) suggest that a variety of methods is necessary as some studies have shown that certain methods may produce conflicting results for the same items (see for example, Kristjansson, Aylesworth & McDowell, 2005). Thus, multiple methods for DIF detection were selected for this study. Due to limitations in the sample size 2- or 3-parameter IRT approaches were not suitable (see McNamara & Roever, 2006).

The two DIF detection procedures chosen as methods for the current study were the standardization procedure (also known as conditional p value) (Dorans & Kulick, 1983) and the Mantel-Haenszel procedure (Dorans, 1989; Mantel & Haenszel, 1959). Both procedures involve a comparison between a 'reference group' and a 'focal group'. The focal group is considered the 'group of interest', and the reference group is the group with whom performance is being compared (Holland & Wainer, 1993, p.xv). The standardization and Mantel-Haenszel procedures also involve matching test-takers on

ability level; and each allows for matching to be performed using an external criterion. The selection of these two procedures reflects the approach taken by Roever (2007) in which both methods used together were found to be complementary, and useful for investigations with relatively small sample sizes (e.g., 250). Similarly, Hambleton (2006, p.186) recommends these two procedures for identifying DIF with limited numbers of test-takers.

V Data collection

1. Test materials

The primary research instrument was the UTESL (University Test of English as a Second Language) listening sub-test. The UTESL is an academic English proficiency test developed by the Language Testing Research Centre at the University of Melbourne, Australia. Each version consists of one lecture of approximately 30 minutes in length (including pauses and instructions), which is divided into four sections, and usually contains between 30 and 40 items. The specifications of the UTESL listening sub-test identify four key components of listening ability that are tapped by the items:

- Summarise main points
- Locate and recall specific information
- Distinguish between main points and supporting detail
- Reorganise information from the lecture to complete a graph, chart or diagram

A wide range of open-ended and fixed-choice task types are used on the UTESL, including gap fills, table completion, short answer questions and multiple choice questions.

2. Speakers

Three speakers, each representing a particular accent group – Japanese-, Mandarin Chinese- (henceforth Chinese) and Australian-English – were selected to re-record existing UTESL materials. This involved a rigorous process which was designed to select speakers according to the criteria that, (1) speakers had equivalent levels of general intelligibility; (2) speakers were not perceived to be unreasonably difficult to understand; and (3) L2 speakers had accents which were identifiably L2 varieties. To achieve these aims, a pool of nine speakers (3 per accent) recorded excerpts of oral stimuli drawn from UTESL listening sub-test scripts. Then, following methods described in Munro and Derwing (1995), a group of 20 listeners from a range of L1 backgrounds completed a set of tasks designed to measure speaker intelligibility (through a transcription task) while also eliciting subjective ratings of comprehensibility and accentedness, and perceptions of accent identification. These measures yielded a range of data from which a decision was made to select three specific speakers – Henry (Australian English accent), Kaori (Japanese accent) and Jun (Chinese accent). Henry, Kaori and Jun were each asked to record one version of an existing UTESL listening sub-test: the 'Food Technology' test (recorded by Henry – Australian English accent), the 'Oldest Old' test (recorded by Jun – Chinese accent), and the 'Sleep' test (recorded by Kaori – Japanese accent). These three tests formed the diverse-accents (DA)-UTESL battery.

3. Language Experience Questionnaire

A Language Experience Questionnaire (LEQ) was designed to collect biographical data from research participants (including participants' first language), and to gauge participants' familiarity with Japanese, Chinese and Australian English accents. Familiarity was measured by self-report with a response on a Likert scale ranging from 1

(not familiar) to 5 (very familiar). These self-reported familiarity ratings were then validated against a range of other measures designed to capture an overall view of participants' exposure to accents of English (see Harding, 2011 for a detailed description of the validation of the LEQ).

4. Listeners

212 participants from a range of Australian universities and English language centres took part in the study. All participants spoke English as a second language, and represented a range of first-language (L1) backgrounds: 70 test-takers with a (Mandarin) Chinese L1 background, 60 with a Japanese L1 background and 82 participants with a range of other L1 backgrounds (see Table 1).

Table 1: Participants by first language (L1)

L1	N
Chinese	70
Japanese	60
Korean	35
Spanish	12
Indonesian	10
Arabic	9
Vietnamese	4
Thai	4
French	3
Turkish	2
Portuguese	1
Hindi	1
West Ambae language	1
Total	212

At the time of data collection, all participants were residing and studying in Australia, and thus could be expected to have had some degree of exposure to Australian English. An analysis of the LEQ revealed that reported familiarity with a Chinese English accent was

significantly higher among Chinese L1 listeners (M = 4.49, SD = 0.79) compared with other L1 listeners [M = 2.14, SD = 1.09; t (180.40) =17.80, p < .001]. Similarly, reported familiarity with a Japanese English accent was significantly higher among Japanese L1 listeners (M = 4.37, SD = 0.94) compared with other L1 groups [M = 2.43, SD = 1.18; t (135.49) = 12.53, p < .001]. These findings suggest that shared-L1 listeners in both instances can also be characterized as highly familiar with their own varieties.

5. Procedure

The DA-UTESL was administered in a series of trials purposefully arranged for data collection. Each trial followed the same procedure: participants received a booklet containing answer papers for all three tests together with the LEQ (included last). Test input was then presented on a CD which included all instructions and pauses. Following the completion of the DA-UTESL, participants filled out the LEQ. The order of tests was reversed for half of the listener population to control for order and fatigue effects.

6. Test quality

Scoring of each test was conducted in accordance with the existing marking guides for the UTESL listening sub-tests. Table 2 shows that all three tests had acceptably high reliability coefficients, and reasonably low estimates of standard error of measurement (SEM):

Table 2: Reliability and SEM by test version

Test	N of items	alpha	SEM
Food Technology (Henry – Australian English accent)	37	.892	2.28
Sleep (Kaori – Japanese accent)	38	.880	2.53
Oldest Old (Jun – Chinese accent)	31	.872	2.03

VI DIF detection procedures

1. Design

Two separate analyses were undertaken to detect DIF in the data: a 'shared-L1 analysis' within the Sleep test (Kaori – Japanese accent), and a 'shared-L1 analysis' within the Oldest Old test (Jun – Chinese accent). In each analysis, the focal group was defined as shared-L1 listeners, and the reference group consisted of all other L1 listeners (see Figure 1).

Figure 1: An overview of DIF detection procedures

	Sleep test: Kaori (Japanese accent)	**Oldest Old test: Jun (Chinese accent)**
Focal group	Japanese L1 background listeners (N=60)	Chinese L1 background listeners (N=70)
Reference group	Other L1 background listeners (N=152)	Other L1 background listeners (N=142)

2. Matching groups

The standardization and Mantel-Haenszel procedures each require groups of test-takers to be matched at ability level. This is most commonly achieved in DIF studies by matching test-takers on their total test (or test component) score, and is known as 'internal matching'. A less common approach is to match groups according to an external criterion measure in the form of a parallel version of the test which is administered to test-takers alongside the test version of interest (see Ferne & Rupp, 2007). This was seen as a useful approach for the current study, especially because a shared-L1 advantage had the potential to be pervasive rather than located in a small number of items, making such an investigation prone to the problems inherent in internal matching (see Elder, 1997). Thus, the Food Technology test version (Henry - Australian English accent) was used as the

external criterion. No purification procedure was carried out on the Food Technology test as the rationale behind its use was that it represented the 'status quo' approach of the UTESL to speaker accent: standard native speaker English. If DIF were then to be found on tests delivered by Jun or Kaori, this would be in relation to candidates' performance on this orthodox, unmodified UTESL which represents an existing, reliable measure of the construct of interest: academic English listening proficiency. Tables 3 and 4 show the descriptive statistics for performance on the Food Technology test across each of the group contrasts.

Table 3: Descriptive statistics for Japanese L1 background listeners and Other L1 background listeners on the Food Technology test

	N	Mean	SD
Japanese L1	60	17.98	6.70
Other L1	152	17.01	7.05

Table 4: Descriptive statistics for Chinese L1 background listeners and Other L1 background listeners on the Food Technology test

	N	Mean	SD
Chinese L1	70	17.14	8.13
Other L1	142	17.35	6.32

To achieve an accurate match (using total test score as a matching variable, whether internal or external), DIF practitioners generally advise that the matching criterion should include as many score levels as the data allows (e.g. Angoff, 1993). However this type of 'thin matching' is not always possible in practice, and is problematic when sample sizes are small. In order to avoid these problems, a 'thick matching' system was employed following Donoghue and Allen (1993) in which total score levels on the matching criterion are pooled. The matching process resulted in a set of six score levels across the 37 marks available on the Food Technology test which fit the spread of scores for the

Chinese L1 background listener group. However, because of the relatively lower sample size of the 'Japanese L1 background listener' group, five score levels were created to be used for analysis of the Sleep test. The score levels and their N sizes are shown in Tables 5 and 6.

Table 5: Score levels and sample sizes by Japanese L1 status

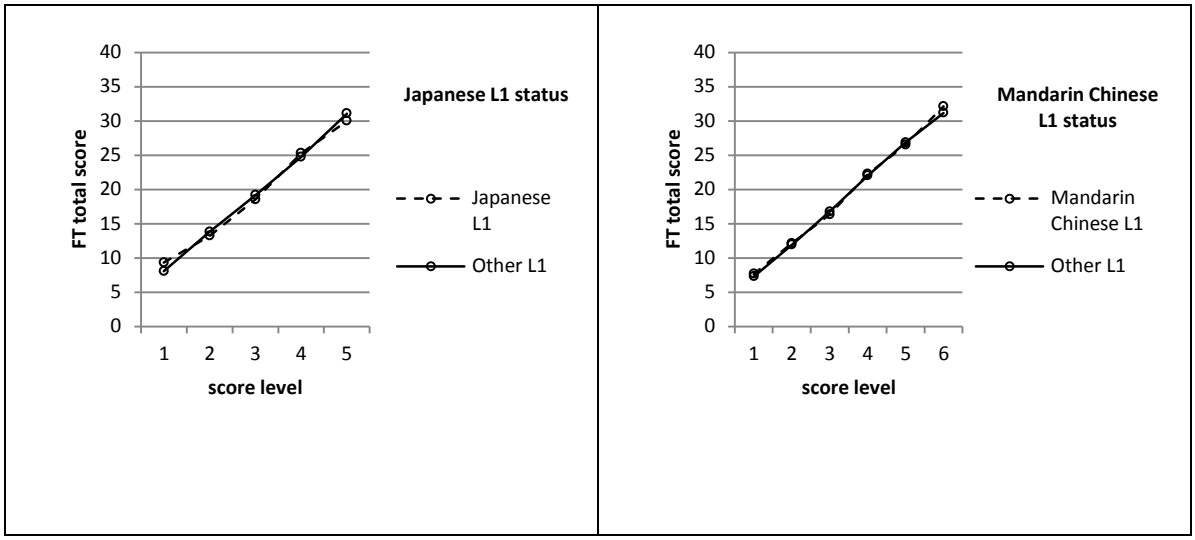
	1 - 10	11 - 16	17 - 21	22 - 27	28-37	Total
Focal group (Japanese L1)	6	22	17	7	8	60
Reference group (Other L1)	30	53	32	24	13	152
Total	36	75	49	31	21	212

Table 6: Score levels and sample sizes by Chinese L1 status

	1 - 9	10 - 14	15 - 19	20 - 24	25 - 29	30 - 37	Total
Focal group (Chinese L1)	15	17	12	12	6	8	70
Reference group (Other L1)	10	39	50	19	18	6	142
Total	25	56	62	31	24	14	212

In order to ensure that groups within these thickly-matched score levels did not differ from each other at each stratified score level, two two-way ANOVAs were conducted with Food Technology total score as the dependent variable, and group membership and score level as independent variables. Results showed that for each contrast there was no significant interaction between group membership and score level at $p < .05$ (see Figure 2). This lends weight to the argument that groups were of matched ability at each step of the stratified score levels.

Figure 2: Focal and reference groups on matching criterion by score level



3. Calculating DIF

The calculation of the standardized p-difference followed the formula provided by Dorans (1989), with focal group N size used as a common weight at each score level:

$$D_{STD} = \sum_{s=1}^{S} N_{fs}\left(P_{fs} - P_{bs}\right) / \sum_{s=1}^{S} N_{fs}$$

Where s = score level (ability level), P_f = observed performance of focal group on item, and P_b = observed performance of the reference group on item (referred to as the 'base group' by Dorans). The resulting figure – the standardized p-difference – ranges from -1 to +1, with each interval of .1 indicating a ten percent advantage for one group over the other on a given item. Dorans and Holland (1993) suggest that values of 0.1 or above and -0.1 or below 'should be examined very carefully', and Roever (2007) considers such values evidence of 'large DIF' (p.179).

The Mantel-Haenszel statistic was calculated for each item through the Crosstabs function of SPSS. In many studies utilising the Mantel-Haenszel procedure, logarithmic

transformations are converted to the delta metric, which is a scale that has a mean of 13 and a standard deviation of 4. This is achieved by multiplying the natural logarithm of the odds ratio by -2.35. The resulting value is known as the Mantel-Haenszel delta difference or MH D-DIF. Conversion to the delta metric allows for easier comparisons between items where, intuitively, items with greater values show stronger DIF in either a negative or positive direction. All results were converted to MH D-DIF for clarity of presentation. Dorans and Holland (1993) suggest that items where MH D-DIF is lower than 1, or where the finding is not statistically significant, should be considered 'negligible DIF' (Type A). Where MH D-DIF is significant and over the value of 1, Dorans and Holland classify the item as 'moderate DIF' (Type B). Type C items – showing 'large DIF' – are those which have an MH D-DIF of above 1.5, and where this value is significantly different from a value of 1. Because of the small sample size in this study, the results below will report all findings where the criteria for Type B DIF are met. Thus the findings include items with at least moderate DIF, and potentially some with large DIF.

VII Results of DIF analyses

1. DIF in Sleep (Kaori – Japanese accent)

The shared-L1 analysis of the Sleep test compared the performance of listeners from Japanese L1 backgrounds (focal group) with listeners from all other L1 backgrounds (reference group). Of the 38 items in the Sleep test, the standardization procedure detected ten items (26%) as having a standardized p-difference of over 0.1, and the Mantel-Haenszel procedure detected seven items (18%) which were significant at $p < .05$, and where MH D-DIF was higher than 1. DIF items were found to advantage both the focal group and the reference group, but with a slight tendency towards the Japanese L1 background listeners as shown in Table 7.

Table 7: Number of DIF items detected in Sleep (38 items) by method

DIF detection method	Focal group advantaged	Reference group advantaged	Total
Standardization	6	4	10
Mantel-Haenszel	4	3	7

There was considerable overlap in these findings; all seven items flagged by Mantel-Haenszel were also flagged by the standardization method. In addition, the standardization method flagged three items that Mantel-Haenszel did not (29, 30 and 31). There were no disagreements between the methods in whether DIF favoured focal or reference groups. Table 8 shows the details and DIF indices of flagged items.

Table 8: Overview of flagged items in Sleep (38 items) and DIF indices

Item	Item type	STD P-DIF	MH D-DIF	*p*	Advantaged group
2	sentence completion	.163	2.026	.017*	focal
6	gap fill	-.209	-2.357	.005*	reference
10	gap fill	.198	2.096	.006*	focal
12	gap fill	.146	1.626	.041*	focal
19	listing	-.131	-1.786	.050*	reference
26	sentence completion	-.149	-2.117	.028*	reference
27	sentence completion	.149	-2.108	.018*	focal
29	multiple choice	-.101	(-1.314)	.124	reference
30	short answer question	.130	(1.426)	.070	focal
31	matching	.137	(1.537)	.061	focal

These findings show that DIF items were somewhat balanced in whether they favoured the reference group or the focal group, though with a slightly higher number of DIF items favouring shared-L1 listeners. This is not uncommon in testing practice, and McNamara and Roever (2006) note that 'DIF favouring the reference group is often balanced out by

DIF favouring the focal group' (p.85). This is not to say that a shared-L1 effect was not present on those items favouring the focal group, but that this effect seems to have been equalised, to a certain extent, by items which favoured the reference group.

2. DIF in the Oldest Old (Jun – Chinese accent)

The shared-L1 analysis of the Oldest Old test compared the performance of Chinese L1 background listeners (focal group) with listeners from all other L1 backgrounds (reference group). This analysis produced markedly different results from the analysis conducted on the Sleep test. Of the 31 items on the Oldest Old, the standardization procedure identified ten items (32%) as having a standardized p-difference of over 0.1, and the Mantel-Haenszel procedure detected eight items (26%) which were significant at $p < .05$, and where MH D-DIF was higher than 1. Of particular interest, DIF items were found to advantage the focal group over the reference group in all but one case (see Table 9).

Table 9: Number of DIF items detected in The Oldest Old (31 items) by method

DIF detection method	Focal group advantaged	Reference group advantaged	Total
Standardization	9	1	10
Mantel-Haenszel	8	0	8

All of the eight items flagged by Mantel-Haenszel were also flagged by the standardization method. In addition, the standardization method flagged two items that that Mantel-Haenszel did not (7 and 24). Table 10 shows the details and DIF indices of flagged items.

Table 10: Overview of flagged items in The Oldest Old (31 items) and DIF indices

Item	Item type	STD P-DIF	MH D-DIF	*p*	Advantaged group
1	short answer question	.235	2.693	.002*	focal
2	short answer question	.166	2.148	.014*	focal
3	short answer question	.219	4.385	.000*	focal
7	listing	.109	(1.488)	.106	focal
8	listing	.166	2.517	.004*	focal
17	information transfer	.155	2.583	.015*	focal
22	table completion	.104	2.688	.034*	focal
23	table completion	.296	3.071	.000*	focal
24	table completion	-.106	(-1.788)	.058	reference
25	gap fill	.193	2.458	.002*	focal

In contrast to the shared-L1 analysis on the Sleep test, the findings presented above show that DIF favoured the focal group on all but one of the items. This suggests that Chinese L1 listeners were clearly advantaged across a range of items on the Oldest Old test.

VIII Discussion

1. Evidence of a shared-L1 advantage

The results of the DIF analysis show that shared-L1 effects were not the same across two tests featuring highly intelligible speakers with L2 accents. On the Sleep test (Japanese accent), DIF was detected in ten items using the standardization procedure and seven items using Mantel-Haenszel, and was fairly evenly balanced for focal and reference groups. By contrast, of the ten items on the Oldest Old test (Chinese accent) which were shown to exhibit DIF by the standardization or Mantel-Haenszel procedures, nine advantaged shared-L1 listeners. However even though a shared-L1 effect seemed clearer on the Oldest Old test, DIF was not shown to be pervasive across all items. These findings, then, reflect many of the preceding studies which have found contradictory evidence for a shared-L1 effect (e.g., Major et al., 2002; Munro et al., 2006), but they do

suggest that, in some circumstances at least, listeners may indeed perform better on a listening test when a speaker shares their L1 background.

Two points of interest in these findings are, firstly, why certain items on the Sleep test were found to favour the reference group, and secondly, why DIF was confined to only a sub-set of items on the Oldest Old test. On the first point, it is possible that the DIF against shared-L1 listeners on four items on the Sleep test may have been a reflection of different levels of (construct-relevant) linguistic knowledge across the two groups ('item impact' according to Zumbo, 2007) rather than any advantage related to accent. For example, item 6 on the Sleep test, which showed DIF in favour of the reference group (Other L1 listeners), required the test-taker to fill a gapped text with the word 'environmental' (which is stated directly in the input). On the Oldest Old test, item 23 required the same term, 'environmental' to be noted down (albeit in a different context), and this item showed DIF in favour of Chinese L1 listeners (who also comprised the largest L1 group within the "Other L1" group in the Sleep test analysis). Taken together, this may suggest that the DIF finding on item 6 of the Sleep test was related to a generally lower awareness of this lexical item for this particular group of Japanese L1 listeners compared with matched-ability counterparts in the reference group (or a higher awareness of this item by the Chinese L1 listeners). The small sample size of the Japanese L1 group would mean that irregularities such as these would be more glaring, and perhaps not indicative of what would be found with a broader population.

On the second point, the reason why DIF was not pervasive across all items on the Oldest Old test appeared to be related to certain characteristics of those items which did show DIF. In a fuller discussion of candidates' responses to DIF items (Harding, 2011) it is shown that the items on the Oldest Old test which showed DIF were likely to be those which focused on bottom-up listening skills, where the linguistic demands of the input

were high, and where the speaker's pronunciation was both characteristic of documented features of a Chinese English variety (according to sources such as Chang, 2001), and also exhibited pronunciations which deviated from core features that have been identified as crucial for lingua franca intelligibility (Jenkins, 2000), such as consonant substitutions and deletions and non-standard nuclear stress. This type of post-hoc analysis of DIF items can only result in hypotheses about why some items showed DIF and others did not. However, it provides a starting place for further empirical research on the influence of task and speaker variables, and suggests that, with a full understanding of the conditions under which a shared-L1 effect is more likely to occur, that such effects may be made 'manageable' in the test development process.

2. Conceptualising a shared-L1 effect as bias

The next step is to assess the degree to which evidence of a clear shared-L1 effect, as observed through findings of DIF, would constitute bias. This is a complex question, and will depend on the purpose of the test, the nature of the target-language use domain, and/or the prominence of the 'ability to cope with accent variation' in the construct definition. For example, it may be the case that the TLU domain of a listening test has been described exhaustively, and that specific L2 accents have been identified as particularly important to include in listening materials. Thus, if the Oldest Old test were to be used in a situation where the ability to listen to a Chinese accent had been identified in the construct specifications, then the variation shown in the DIF findings should not be considered bias, but can be considered variation which is directly construct-relevant. On the other hand, if the ability to listen to L2 accents is not considered part of the construct under test, then a DIF finding – such as the one observed on the Oldest Old test – is

concomitant to a finding of bias, as performance between the two equally matched groups varies according to a factor which is not construct-relevant.

A more difficult (and perhaps more likely) situation exists if there is a desire to operationalise a construct of 'ability to cope with accent variation (including L2 accents)' within a listening test, but where no specific target accent has been (or can be) identified in the TLU domain. In this situation, the link between the DIF finding and the question of bias is less clear as DIF might be understood as indicating, not construct-irrelevant variance, but rather construct under-representation for one unique group: the shared-L1 listeners. That is, the ability to cope with accent variation is posed by the introduction of a particular L2 speaker for all but the group of listeners who share the speaker's L1. The test may still be thought of as biased, but the problem is not solved by eliminating the source of variance, because the result would be an impoverished construct.

IX Implications

This research adds to a number of studies which have shown that a shared-L1 effect in listening comprehension is at the very least possible, and in certain circumstances, clear. At the same time, there is an increasing need for language tests to grapple with the sociolinguistic reality of English language domains, and this includes an acknowledgement that learners will encounter and need to deal with a range of accents, including L2 varieties, and that this should be reflected in listening test constructs. Ultimately, then, the question of whether or not to introduce L2 speech in listening assessment becomes a policy issue at the thinking stage of test construction. If it is decided that, for the purposes of the test, the ability to process L2 varieties is not part of the construct, then it is clear that the use of L2 speakers may result in construct-irrelevant variance (although it is not clear what sort of domain could be described in the early 21st

century where a listener would be required to listen only to native speaker varieties). If, on the other hand, it is decided that the ability to deal with L2 varieties is directly construct relevant, then a decision needs to be made about how best to incorporate this into the modelling of the construct in a way where the listening demands associated with an L2 accent are experienced equally (to the greatest degree possible) across all L1 groups.

In balancing these competing demands, three approaches may be considered:

(1) Conduct a thorough needs analysis to identify which accents are particularly salient in a given context, and only include those accents in listening test materials. The advantages of this approach would be that, once a set of target accents has been identified, a shared-L1 effect will not threaten a validity argument. However this procedure would be difficult to implement, particularly for large-scale, international tests which hope to generalise scores across multiple domains.

(2) Attempt to 'manage' the impact of DIF through the use of highly intelligible L2 speakers (and perhaps attempt to control those task and input factors which have been hypothesised to contribute to the likelihood of DIF). This approach would reduce the potential for unequal construct representation across groups, and would perhaps have greater face validity with stakeholders (it is similar to the current approach adopted on Cambridge ESOL general English exams where non-native speaker voices are heard which 'approximate to the norms of native speaker accents' [Cambridge ESOL, n.d.]). However it would not necessarily broaden the construct as the input would have an L2 'flavour' only. Moreover, in limiting the range of task-types and the linguistic difficulty of input there is a risk

of under-representing the listening construct in other ways which may not be desirable.

(3) Following the recommendations of Buck (2001), attempt to balance the impact of DIF by using several speakers across a range of tasks. Through this approach, the challenges represented by listening to L2 accents may be retained and tested across all listener groups, although some listener groups will never hear their own accents and so materials may be perceived as unfair. Alternatively, an accent may be chosen which is equally unfamiliar to all; however this will only be possible in testing contexts where background information about the candidature is known in advance, or where the test-taker population is homogeneous.

These are tentative solutions, and clearly the practical constraints of particular testing situations will make some approaches more feasible than others.

Further research is also required to achieve a deeper understanding of the nature of a shared-L1 effect, and the ways in which listeners deal with L2 accents in a testing context. Firstly, studies replicating the DIF methods used here with different accents, in different testing contexts, and with greater numbers of listeners are required to see under which conditions a shared-L1 effects holds. Secondly, the influence of task and pronunciation variables on a shared-L1 effect needs to be investigated systematically in order to provide a better understanding of the circumstances under which accent-related DIF is most likely to occur. Thirdly, the element of the listening construct relating to 'ability to deal with an accent' needs to be understood more fully through introspective methods which seek to understand how test-takers deal with accent in the context of a listening test. Finally, the degree to which short-term adaptation to accent might occur during the course of second language listening assessment has not been established. An

investigation of this phenomenon would not only have implications for an understanding of the ways in which test-takers might engage in 'perceptual learning' throughout the listening test event, but would deepen current conceptualisations of the construct of listening ability more generally.

Acknowledgement

This paper represents a revised version of research from my PhD thesis, which has been published as a book through Peter Lang (see Harding, 2011).

References

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-24). Hillsdale: Lawrence Erlbaum Associates.

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

Best, C. T. (1995). A direct-realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-cultural research* (pp. 171-204). Baltimore: York Press.

Bradlow, A., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition, 106* (2), 707-729.

Buck, G. (2001). *Assessing listening*. Cambridge: Cambridge University Press.

Cambridge ESOL (n.d.). *Candidate support: CPE – the papers*. Retrieved April 1, 2011 from http://www.candidates.cambridgeesol.org/cs/Help_with_exams/General_English/CPE/Papers?paper=Listening&panel=faqs

Cambridge ESOL (2008). *IELTS teaching resource*. Retrieved March 1, 2008 from www.cambridgeesol.org/teach/ielts/listening/aboutthepaper/overview.htm

Canagarajah, S. (2006). Changing communicative needs, revised assessment objectives: Testing English as an international language. *Language Assessment Quarterly, 3* (3), 229-242.

Chang, J. (2001). Chinese speakers. In M. Swan and B. Smith (Eds.), *Learner English: A teacher's guide to interference and other problems* (pp. 310-324). Cambridge: Cambridge University Press.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *Journal of the Acoustical Society of America, 116* (6), 3647-3658.

Davies, A., Hamp-Lyons, L., & Kemp, C. (2003). Whose norms? International proficiency tests in English. *World Englishes, 22* (4), 571-584.

Donoghue, J. R., & Allen, N. L. (1993). Thin versus thick matching in the Mantel-Haenszel procedure for detecting DIF. *Journal of Educational Statistics, 18* (2), 131-154.

Dorans N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. *Applied Measurement in Education, 2* (3), 217-233.

Dorans, N. J., & Kulick, E. (1983). *Assessing unexpected diferential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of the standardization approach* (ETS Research Rep. RR-83-9). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale: Lawrence Erlbaum Associates.

Elder, C. (1997). *The background speaker as learner of Italian, Modern Greek & Chinese: Implications for foreign language assessment*. Unpublished doctoral dissertation, The University of Melbourne.

ETS [Educational Testing Service] (2005). *TOEFL iBT Tips: How to prepare for the next generation TOEFL test and communicate with comfort*. Princeton, NJ: ETS.

Ferne, T., & Rupp, A. A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges and recommendations. *Language Assessment Quarterly, 4* (2), 113-148.

Flowerdew, J. (1994). Research of relevance to second language lecture comprehension – an overview. In J. Flowerdew (Ed.), *Academic listening* (pp. 7-29). New York: Cambridge University Press.

Hambleton, R. K. (2006). Good practices for identifying differential item functioning. *Medical Care, 44* (11), 182-188.

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

Holland, P. W., & Wainer, H. (Eds.) (1993). *Differential item functioning*. Hillsdale: Lawrence Erlbaum Associates.

IELTS [International English Language Testing System] (2007). *IELTS Handbook.* British Council, IDP: IELTS Australia, University of Cambridge ESOL Examinations.

Jenkins, J. (2000). *The phonology of English as an international language*. Oxford: Oxford University Press.

Jenkins, J. (2006). The spread of EIL: A testing time for testers. *ELT Journal, 60* (1), 42-50.

Kristjansson, E., Aylesworth, R., & McDowell, I. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement, 65*, 935-953.

Llurda, E. (2004). Non-native-speaker teachers and English as an International Language. *International Journal of Applied Linguistics, 14* (3), 314-323.

Lowenberg, P. (2002). Assessing English proficiency in the Expanding Circle. *World Englishes*, 21, 3: 431-435.

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.

Mantel, N. & Haenszel, W. (1959) Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, **22**, 719-748.

McNamara, T. F., & Roever, C. (2006). *Language testing: The social dimension*. Oxford, England: Basil Blackwell.

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.

Ortmeyer, C., & Boyle, J. P. (1985). The effect of accent differences on comprehension. *RELC Journal, 16* (2), 48-53.

Roever, C. (2007). DIF in the assessment of second language pragmatics. *Language Assessment Quarterly, 4* (2), 165-189.

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.

Smith, L.E., & Rafiqzad, K. (1979). English for cross-cultural communication: The question of intelligibility. *TESOL Quarterly, 13* (3), 371-380.

Stibbard, R. M., & Lee, J. (2006). Evidence against the mismatched interlanguage speech intelligibility benefit hypothesis. *Journal of the Acoustical Society of America, 120* (1), 433-442.

Strange, W. (1995). Cross-language studies of speech perception: a historical overview. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-cultural research* (pp. 3-45). Baltimore: York Press.

Tauroza, S., & Luk, J. (1997). Accent and second language listening comprehension. *RELC Journal, 28* (1), 54-71.

Taylor, L. (2006). The changing landscape of English: implications for language assessment. *ELT Journal, 60* (1), 51-60.

Yule, G., Wetzel, S., & Kennedy, L. (1990). Listening perception accuracy of ESL learners as a function of speaker L1. *TESOL Quarterly, 24* (3), 519-523.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4* (2), 223-233.