The Intl. Journal of Listening, 26: 40–60, 2012

ISSN: 1090-4018 print / 1932-586X online
DOI: 10.1080/10904018.2012.639649

Routledge
Taylor & Francis Group

Differential Item Functioning in While-Listening Performance Tests: The Case of the International English Language Testing System (IELTS) Listening Module

Vahid Aryadoust

English Language and Literature Academic Group
National Institute of Education, Nanyang Technological University

This article investigates a version of the International English Language Testing System (IELTS) listening test for evidence of differential item functioning (DIF) based on gender, nationality, age, and degree of previous exposure to the test. Overall, the listening construct was found to be underrepresented, which is probably an important cause of the observed lack of significant correlation between awarded scores on while-listening performance (WLP) tests and subsequent academic performance. Some short answer items were biased toward higher-ability subgroups, likely due to those test takers' higher ability to apply what they had understood. Finally, some multiple-choice questions (MCQs) with few options likely encouraged attempts at lucky guesses, particularly among low-ability people who had received training in test-taking strategies. Implications for listening assessment and language education are discussed.

Listening tests, which require test takers to demonstrate comprehension of spoken language, are widely used in second language assessment. However, listening tests are understood not to be "pure" evaluation tools in that they require test takers to apply skills other than listening, such as reading and writing (Bejar, Douglas, Jamieson, Nissan, & Turner, 2000, p. 59). This complexity renders all the more important the minimization of construct-irrelevant influences, such as the test's response format and test takers' background, topic familiarity, and test-taking strategies (Bachman, 1990; Haladyna & Downing, 2004).

A number of studies have examined the constituent subskills that lead to successful performance in listening tests. Two lines of research emerge in the pertinent literature. The first comprises studies investigating how listening comprehension subskills account for variation in scores (Bodie, Worthington, & Fitch-Hauser, 2011; Richards, 1983; Shohamy & Inbar, 1991; Tsui & Fullilove, 1998; Wagner, 2002, 2004, 2010; Goh & Aryadoust, 2010). Most directly, Aryadoust and Goh (2010) identified several major subskills of successful listening performance, beginning with the ability to use schemata to relate an oral message to the real world and including the ability to identify the topic, identify supporting examples, and make inferences and draw conclusions.

Correspondence concerning this article should be addressed to Vahid Aryadoust, English Language and Literature Academic Group, National Institute of Education, Nanyang Technological University, 1 Nanyang Walk, Singapore 637616. E-mail: vahidaryadoust@gmail.com or Ng0666907m@stdmail.nie.edu.sg

The second line of research comprises studies exploring the effects of various test taker and environmental attributes on performance (Aryadoust, Goh, & Lee, 2011; Bejar et al., 2000; Brindley, 1998; Hansen & Jensen, 1994; Goh & Aryadoust, 2010). Powers (1985, pp. 23–25) categorized the variables affecting listening comprehension in academic contexts as "stimulus-related," "speaker-related," or "context-related"; that is, pertaining to the message itself (e.g., its vocabulary range, vagueness, and "visual cues"), to the manner in which the message is delivered (e.g., the speed, stress, intonation, and rhythm), or to the features of the context in which the exchange takes place (e.g., note-taking and distractions during communication). Among these, only variables targeting test takers' trait levels in the various listening comprehension subskills are desirable. Test taker, speaker, and environmental variables not directly connected to the listening construct must be removed.

One useful distinction in listening tests is between postlistening and while-listening performance tests. In postlistening performance (PLP) tests, students listen to oral stimuli, take notes, and then answer test questions (Chen & Henning, 1985; Elder, 1996; Muraki, 1999). Answering questions in PLP tests is a postcomprehension activity. Two of the best-recognized PLP tests are the Test of English as a Foreign Language (TOEFL®) and the Michigan Language Assessment Battery (MELAB) listening tests. In while-listening performance (WLP) tests, test takers read and answer test items *while* they listen to oral stimuli, and thus engage in the following simultaneous activities: (a) read test items, (b) listen to the oral text, (c) write or choose the answer, and (d) follow the oral text to move to the next test item (Goh & Aryadoust, 2010). The best-known WLP tests are designed by the University of Cambridge ESOL (English for Speakers of Other Languages) Examination Syndicate, and include the International English Language Testing System (IELTS™) and the Certificate in Advanced English (CAE) tests.

The simultaneity of WLP test performance raises a number of important theoretical points not yet fully addressed in the extant literature. First, since these tests do not differentiate listening comprehension from the ability to subsequently apply the comprehended input (see Field, 2009), they can mislead test developers and score users about the listening construct and uses of test scores (Dunkel, Henning, & Chaudron, 1993). WLP tests also seem to confuse precursor subskills with listening comprehension, failing to address vital facets of comprehension. For example, some items in WLP tests limit their focus to phoneme and word juncture recognition skills (Aryadoust, 2011a, b; Aryadoust & Goh, 2010). These basic skills do facilitate comprehension; however, theories of listening comprehension distinguish between the two (Dunkel et al., 1993). Finally, because WLP test takers' simultaneous exposure to oral and written inputs precludes note taking, it is likely that test takers who fall behind the stream of written/oral input miss some items not necessarily because of limited listening skills but rather because of limited reading skills, memory span (Hildyard & Olson, 1978), test-taking strategies (Bachman, 1990), test wiseness (Bachman, 1990; Kunnan, 1995), or other constraining influences (Field, 2009).

A number of researchers have studied the external relationships between WLP test scores and the academic performance of test takers (Cotton & Conrow, 1998; Ingram & Bayliss, 2007; Kerstjen & Nery, 2000; Merrylees, 2003) but only recently have published studies reported on the constituent structure of WLP tests and their interaction with test takers' cognitive processes and strategies (Badger & Yan, 2009; Coleman & Heap, 1998; Field, 2009; Geranpayeh & Kunnan, 2007). For example, Field (2009) argued that WLP test takers have a higher chance of a better performance if they are exposed to test inputs in a "freer" condition where the restrictive circumstances of test taking, item formats, and the needs for concurrent listen-read-write are removed.

Field (2009) showed that these test- and examinee-related factors pose serious threats to the validity of the lecture comprehension section of the IELTS listening test.

At present, the literature on WLP testing is primarily constrained by its lack of differential item functioning (DIF) analysis, emphasized by the *Standards for Educational and Psychological Testing* (American Educational Research Association/American Psychological Association/National Council on Measurement in Education [AERA/APA/NCME], 1999) as an important procedure for evaluating the validity argument of a test. DIF analysis is a method of investigating whether test items function differently (i.e., award significantly different scores) for different subgroup of test takers (Aryadoust et al., 2011a, b ; Ferne & Rupp, 2007; Tennant & Pallant, 2007). Several studies have investigated DIF in second-language listening tests for different subsamples and from different perspectives (Abbott, 2007; Breland, Lee, Najarian, & Muraki, 2004; Conoley, 2004); however, within the limited literature on WLP tests, DIF analysis is the least-researched area (Geranpayeh & Kunnan, 2007) in part because the research policies of the University of Cambridge seem to prioritize reliability analysis over other issues in WLP tests (the Web site of IELTS, n.d.a.). In the absence of such studies, the test users and researcher are left to presume that the tests are fair and test items do not favor any group of test takers.

Banks (1999) and Geranpayeh (2001) argued that there is a need to investigate age- and nationality-related DIF in the WLP tests of the University of Cambridge ESOL Examination Syndicate. More recently, Geranpayeh and Kunnan (2007) argued that the WLP tests developed by the University of Cambridge ESOL Examination Syndicate would likely display DIF for such variables as nationality background, age, and gender. They stated that "there has been a shift in the traditional test population, where test takers of many age [nationality and gender] groups are taking these cognitively challenging tests" (Geranpayeh & Kunnan, 2007, p. 193). This change in the attributes of the test taking population introduces new variables which can potentially affect test fairness (*Standards* [AERA/APA/NCME], 1999). For example, certain item characteristics variables (e.g., passage topics, item location and content, vocabulary) which are sensitive to gender, age, or nationality differences can cause certain items to function differently across these groups (Ferne & Rupp, 2007; Kunnan, 1995).

With these considerations in mind, a major objective of this study was to investigate the interaction between item functioning and four test taker attributes: age, nationality, gender, and previous exposure to WLP tests. All four attributes were predicted to cause significant DIF. The directionality of DIF was not generally predicted, although a high degree of previous exposure to the test was predicted to confer an advantage.

METHOD

Participants

Participants in this study were 209 multinational students with undergraduate and graduate education backgrounds. The study was conducted in three venues: the British Council in Malaysia, the National Institute of Education (NIE) of Singapore, and Multilingual Incorporated in the Philippines. Participants were Iranian (49.8%), Chinese (25.8%), Malay (12.4%), and others (12%)—mainly from Arab states of the Persian Gulf—and between the ages of 16 and 45 ($M = 26.7$; $SD = 5.49$). Participants were informed that their participation in the study was voluntary, that their participation and answers would be kept confidential, and that they would be

provided with complete feedback on their performance. Participants filled out consent forms prior to participating in the study.

Materials

Participants were given a version of the IELTS listening test, a WLP test administered globally by the three bodies responsible for IELTS: the British Council, University of Cambridge ESOL Examinations, and IDP Australia. Like other Cambridge ESOL exams, IELTS is developed through the seven-stage Cambridge ESOL Question Paper Production Cycle (CEQPPC), which is a rigorous process of collecting evidence for the validity of the uses and interpretation of the test scores.

The IELTS listening module comprises four sections that evaluate test takers' ability to understand spoken English in different contexts. Sections 1 and 2 evaluate comprehension of everyday conversation, and sections 3 and 4 assess comprehension of academic discourse. Section 1 exposes listeners to a conversation and tests their understanding of specific and factual information; section 2 has the same assessment objective as section 1, but the stimulus is a short radio talk or an excerpt from a monologue; section 3 is "a conversation which involves negotiation of meaning [and] listening for specific information, attitudes, and speakers' opinions" in an academic context (The University of Cambridge ESOL Examination Syndicate, n.d.); and section 4 has the same assessment objective as section 3, but the stimulus is an academic monologue.

Each of the four sections has 10 test items, which fall into seven types: (a) forms/notes/table/flow-chart/summary completion, (b) multiple-choice questions (MCQs), (c) short-answer questions, (d) sentence completion, (e) labeling a diagram/plan/map, (f) classifying, and (g) matching. MCQs and some other item types do not require students to generate the correct answer on their own, but sentence completion and other short production items do. In all cases, test takers mark their answers concurrently while listening to the stimuli.

Because analysis of a "live" listening test is impossible for confidentiality reasons, study participants were given a listening test from *Official IELTS Practice Materials* (www.IELTS.org, 2007). This test was created using the seven-stage Cambridge ESOL CEQPPC process, the same process by which all Cambridge-developed tests are constructed, and is designed to present test takers with exactly the same types of stimuli as a "live" test. The training manual in which it appears "is the only IELTS training book endorsed by the IELTS partners" (the Web site of IELTS, n.d.b.), those partners being the British Council, IDP IELTS Australia, and Cambridge ESOL.

The administered test included five out of the seven test item types: table completion, MCQ, short answer items, sentence completion, and labeling. Not included were classifying and matching test items.

Procedures

As noted above, the test was administered to participants in Singapore, Malaysia, and the Philippines. Data from Singapore and Malaysia were collected by the researcher, and data from the Philippines were collected by a trained colleague.

Each test administration was conducted in accordance with standard IELTS testing rules and procedures and under similar conditions to minimize the effect of extraneous variables. Participants were given a short introduction into the IELTS listening test to familiarize them

with the test structure and an abridged 10-item sample test (taken from *IELTS Practice Test Plus* by Jakeman & McDowell, 2001) to allow them to practice the rules they had learned. Next, participants were given the 40-minute IELTS listening test. In accordance with IELTS procedure, participants were allotted time to read ahead in the question booklet and 10 minutes to transfer their answers to answer sheets. Following the test, participants were asked to provide their demographic information.

Participants received a hard copy of the test after the test session, and interested participants received their results via e-mail, including raw scores, IELTS band score equivalents, problem areas, and suggestions for improvement.

Data Analysis

Qualitative Item Analysis

The researcher and two item reviewers undertook a qualitative investigation of the test items. The purpose of this analysis was twofold: to investigate the test items' structure, content, and form and to provide qualitative evidence regarding the dimensionality of the test (Kim & Jang, 2009). Dimensionality analysis, a prerequisite for latent trait data analysis tools, is concerned with the range and scope of subskills that a test taker needs to answer test items. For example, a listening test might be expected to test language learners' ability to use lexis and syntax effectively in understanding explicit meaning and making inferences. If test takers must use other skills beyond this range, such as reading or writing, these additional skills form off-target dimensions in addition to the major listening dimension, and might complicate test score use and interpretation.

For each test item, the researcher and reviewers noted impressions of (a) the subskills required to answer the item, (b) listening strategies that test takers might use to answer the item, (c) difficulties that test takers would likely encounter answering the item, and (d) construct-irrelevant factors that could potentially affect performance on the item.

We found most of our views on each test item complimentary, and almost all views were kept in the finalized content analysis report. In several instances, our judgments varied in their level of "specificity" (Kim & Jang, 2009, p. 838). For example, one reviewer suggested that test items 1 and 2 required the general ability to connect oral stimuli to a visual map, whereas another suggested that the items require that the test taker connect comprehension of specific information and the ability to process visuals. We followed Kim and Jang (2009, p. 839) in attempting to create a list of necessary subskills that were "conceptually distinguishable" and identifiable in "a minimal number of items."

The Rasch Model

The principal analysis undertaken in the study is a Rasch model-based analysis of dimensionality and DIF in the test data. Calculations were made using WINSTEPS computer program, Version 3.67 (Linacre, 2009a).

The Rasch model is a data analysis instrument that generates multi-item interval scales by placing test items (items) and test takers (persons) on a continuum on which the location of items and persons corresponds to their difficulty and ability measures, respectively. The Rasch model estimates item difficulty measures based on the proportion of test takers answering the item correctly and person ability measures based on the number of items answered correctly.

It can then estimate the probability of a particular test taker answering a particular test item correctly as a function of the relevant ability and difficulty measures.

Contrary to classical testing theory (CTT), every person and item is given a measurement error index estimated according to their compliance with the expectations of the model. In addition, CTT lacks an item-by-person continuum representing a basically unidimensional latent trait and additivity of the scaled data (Prieto, Alonso, & Lamarca, 2003). By contrast, the Rasch model enables the construction of an additive, unidimensional scale and the easy examination of item and person locations. The Rasch model produces a ruler-like device to compare person and item locations. The ruler is an interval scale with logits units (log odds units), with a unidimensional line separating persons and items.

The properties of the Rasch model apply to the extent that the data fit the model. If the data do not fit the model, likely reflecting contaminating secondary dimensions, then Rasch-based item and person measures are likely to be inaccurate. For example, a low-ability individual who answers a number of difficult items correctly is likely to have guessed, or to have benefited from factors other than the ability under assessment (Linacre, 2009b). Persons and items not satisfying the model's predictions are identified as *misfitting* the model.

To measure fit to the Rasch model, two kinds of mean square (MNSQ) fit indices were computed: infit and outfit (Linacre, 2002), both attained by dividing the chi-square values by their degree of freedom (Linacre & Wright, 1994). Infit MNSQ is a *t*-standardized information-weighted statistic sensitive to inliers, and typically produces more useful information about the measurement than outfit MNSQ, a *t*-standardized unweighted statistic sensitive to outliers. (Weighted statistics reduce the impact of "less informative, low variance, off-target responses" [Wright & Masters, 1982, p. 100].) The expected MNSQ value is 1.0, so, for example, a value of 1.2 has 20% "noise" or unmodeled variance more than expected by the model; it can increase the standard error of measurement (Smith, 1996; Wright & Linacre, 1994). Linacre (2009b) regards fit indices between 0.5 and 1.5 as an acceptable range.

The Rasch model for dichotomous data is similar to one-parameter logistic item response theory (1PL IRT). However, there are a few features that make the Rasch model unique. First, the Rasch model parameterizes each individual for estimating item difficulty, whereas 1PL parameterizes the entire sample by the mean score and standard deviation. Second, the Rasch model provides "local" fit statistics, which are diagnostic (Linacre, 2005). They make it possible to examine the performance of every individual and functionality of each item, but the 1PL merely provides a "global" fit to retain or reject the entire model. Third, the Rasch model does not require data to approximate a normal distribution.

Test of Dimensionality

Dimensionality analysis attempts to ascertain whether test items target the same latent trait. If they do, the test is said to be unidimensional. Unidimensionality is a necessary precondition to Rasch-based DIF analysis (see Linacre, 2009b, for a discussion), so a Rasch-based dimensionality analysis of the data precedes the DIF analysis in this study (Aryadoust et al., 2011).

The main analytical tool used to investigate the dimensionality of the test data is principal component analysis of linearized Rasch residuals (PCAR) (which differs from principal component analysis (PCA) of raw data). Rasch fit indices can lead the researcher to erratic patterns in data, but when items or subskills are significantly correlated their power to detect multidimensionality

decreases (Linacre, 1998; Smith, 1996). In contrast, the analysis of linearized Rasch residuals left after extracting the Rasch dimension from the data is an effective method of investigating dimensionality (Linacre, 2009b). It helps compare the observed variance of the data to the "variance components expected for these data if they exactly fit the Rasch model" (Linacre, 2009b, p. 280).

The results of PCAR are illustrated as a map (i.e., *Wright map*) by WINSTEPS. If no substantive pattern appears in item residuals, items scatter in different regions of the map without clustering in either its positive or negative loading regions (Aryadoust et al., 2011). However, substantive dimensions form clusters of items called "contrasts," which form in opposing regions of the plot according to their loading values. Contrasts have positive and negative loading patterns, "reflecting opposing response patterns across items by persons" (Linacre, 2009b, p. 433).

If the observed contrasts explain very little of the observed variance in the data, then the data are said to be unidimensional and meet the requirements of the Rasch model. However, if item residuals form one or more secondary contrasts that explain a significant amount of observed variance, the data are likely to be contaminated with another latent trait, most likely an unmodeled and construct-irrelevant factor. In this case, the data are said to be multidimensional.

Following PCAR of a dataset, Linacre (2009b) and Raîche (2005) suggested that researchers simulate a second, unidimensional data set containing the same number of people and items and with the same person and item reliability indices as the actual data set, and compare its PCAR results to that in the actual data. If the two PCAR results approximate each other, this is further evidence of unidimensionality.

Differential Item Functioning Investigation

DIF analysis investigates the appropriateness and fairness of a test at the item level. If two subgroups of test takers have similar ability levels but markedly different success rates on some items, DIF exists and contaminates test scores (Tennant & Pallant, 2007). The null hypothesis in Rasch-based DIF studies is that, for each test item, there is no interaction between grouping variables and item function, and all subgroups' local difficulty estimates are the same within error (Linacre, 2009b). This hypothesis is rejected if the *t*-test shows that the performance mean indices of different subgroups are statistically different on some items.

DIF can take two forms: uniform and nonuniform. Uniform DIF (UDIF) holds when two subgroups of test takers, such as males and females, perform differently on a test item, by an amount that remains constant across all test taker ability levels. Because the difference in performance remains constant, the item characteristics curves (ICC) of the two subgroups have identical slopes and do not intersect. Nonuniform DIF (NUDIF) holds when the difference in performance varies with ability level. Because their slopes differ in different regions, the two subgroups' ICC curves intersect, and the question favors different subgroups at different ability levels.

The major aim of DIF analysis is to find statistically significant DIF measures with substantive magnitudes. If an item calibrates slightly lower for a group with a negligible standard error, this may be accidental and have no repercussions for the measurement. However, if the DIF is substantive (greater than 0.5 logits), further investigation is necessary. For this purpose, the groups should be further split into subgroups to see whether the DIF is still present in the item. This stage may produce a nonuniform DIF (NUDIF), and once it is evident that the DIF persists, we have evidence regarding systematic actual DIF (Du, 1995).

RESULTS

Descriptive Statistics

Table 1 displays descriptive statistics for the test data, computed on SPSS for Windows, Version 17. A few items had large kurtosis and skewness values, likely because their difficulty measures made one or the other of their distribution tails heavy (see Bachman, 2004). Because the data are dichotomous, an easy or difficult test item implies a lot of 1.00 or 0.00 (correct or incorrect) cases, respectively. This makes the shape of the data asymmetric and skewed toward the side with a lower frequency, so the tail of the distribution becomes very long and the skewness value high.

Analysis of Test Content

Through iterative content analysis, the researcher and reviewers identified the following major processes and subskills as playing significant roles in answering the test items: test takers use their (a) linguistic repertoire (e.g., grammar, vocabulary, syntax), (b) world knowledge sources (schemata), (c) ability to paraphrase, and (d) ability to understand specific factual information such as names and numbers. Test takers who understand the oral stimuli attempt to integrate them with the written test items. In so doing, they integrate their reading ability, visual skills, and short-term memory span. Finally, test takers are likely to attempt lucky guesses if they cannot answer an item. Each of these major component skills and processes is observed in a few items.

More importantly, the reviewers and researcher reached a consensus that the test items did *not* evaluate inference making or the comprehension of implicitly articulated information, which agrees with Geranpayeh and Taylor's (2008) assessment of the IELTS listening test. The Appendix provides a description of the identified sub-skills and corresponding test items.

Rasch Analysis

Table 1 presents item difficulty and MNSQ statistics. The table's "Measure" column expresses item difficulty in logits. Items 8 and 9 were the easiest test items (both with difficulties of -3.03 logits), and item 38 was the most difficult (2.97 logits). Erratic outfit MNSQ statistics were observed in items 3, 8, 9, 14, and 19, but all infit MNSQ statistics fell within the acceptable range from 0.5 to 1.5 (Linacre, 2009b). Items 3, 8, 9, and 19 had low point-measure correlation indices, indicating the presence of unexpected variance (noise) in the response string.

Dimensionality Analysis

PCAR showed that the Rasch dimension explained 43.5% of observed variance in the data, which is extremely close to the Rasch model prediction of 43.4%, indicating that the computation of the Rasch difficulty measures was successful (Linacre, 2009b).

Figure 1 illustrates loading patterns of items on the first hypothesized contrast in the linearized residuals. Items landing above the zero-loading region (shown as a dotted line) are represented by capital letters, and items landing below this region are represented by small letters. The test items

TABLE 1
Descriptive Statistics, Difficulty Measures, Fit Indices, and Item Types

Items	*Mean*	SD	*Skewness*	*Kurtosis*	*Measure*	*Model SE*	*Infit MNSQ*	*Outfit MNSQ*	*Item format*
1	0.78	0.412	−1.395	−0.054	−2.09	0.19	0.87	0.64	LM
2	0.76	0.431	−1.201	−0.564	−1.88	0.18	0.95	0.88	LM
3	0.55	0.499	−0.203	−1.978	2.56	0.15	1.35	1.47	MCQ
4	0.50	0.501	0.010	−2.019	−0.38	0.16	1.01	0.99	MCQ
5	0.50	0.501	0.010	−2.019	−0.38	0.16	1.31	1.48	MCQ
6	0.24	0.428	1.231	−0.489	1.21	0.19	1.14	1.15	MCQ[a]
7	0.61	0.489	−0.444	−1.820	−0.99	0.16	0.81	0.79	TC
8	0.87	0.336	−2.227	2.989	−2.83	0.22	1.12	2.09	TC
9	0.87	0.336	−2.227	2.989	−2.83	0.22	1.17	2.68	TC
10	0.51	0.501	−0.048	−2.017	−0.46	0.16	0.97	0.93	TC
11	0.39	0.489	0.444	−1.820	0.21	0.17	1.24	1.25	MCQ
12	0.74	0.441	−1.083	−0.834	−1.76	0.18	1.06	0.93	MCQ
13	0.84	0.366	−1.890	1.587	−2.55	0.21	0.96	0.77	MCQ
14	0.78	0.415	−1.361	−0.149	−2.05	0.19	0.97	2.02	MCQ[b]
15	0.65	0.479	−0.615	−1.638	−1.20	0.17	0.99	1.08	MCQ[b]
16	0.64	0.481	−0.593	−1.665	−1.18	0.17	1.18	1.46	TC
17	0.43	0.496	0.282	−1.939	−0.01	0.16	1.01	0.98	TC
18	0.49	0.501	0.029	−2.019	−0.36	0.16	0.91	0.87	TC
19	0.67	0.473	−0.705	−1.518	−1.32	0.17	1.37	2.4	TC
20	0.37	0.484	0.549	−1.715	0.35	0.17	0.92	0.85	TC
21	0.31	0.462	0.847	−1.295	0.74	0.18	0.83	0.73	SC
22	0.47	0.500	0.126	−2.004	−0.22	0.16	0.92	0.86	SC
23	0.29	0.456	0.922	−1.161	0.84	0.18	0.75	0.58	MCQ
24	0.13	0.341	2.165	2.712	2.16	0.23	0.98	0.78	SC
25	0.47	0.500	0.126	−2.004	−0.22	0.16	1.00	0.98	SC
26	0.21	0.405	1.466	0.152	1.48	0.20	0.97	0.83	SC
27	0.16	0.366	1.890	1.587	1.91	0.22	1.01	1.13	SC
28	0.58	0.495	−0.322	−1.915	−0.83	0.16	0.84	0.74	SC
29	0.26	0.441	1.083	−0.834	1.03	0.18	0.73	0.54	SC
30	0.45	0.499	0.203	−1.978	−0.12	0.16	0.85	0.74	SC
31	0.47	0.501	0.106	−2.008	−0.25	0.16	1.33	1.38	SC
32	0.25	0.433	1.171	−0.636	1.14	0.19	0.86	0.75	SC
33	0.45	0.499	0.184	−1.985	−0.14	0.16	0.96	0.85	SC
34	0.22	0.419	1.327	−0.240	1.32	0.19	0.89	1.05	TC
35	0.37	0.484	0.549	−1.715	0.35	0.17	0.77	0.71	TC
36	0.11	0.308	2.591	2.759	2.51	0.25	0.96	0.69	TC
37	0.52	0.501	−0.068	−2.015	−0.49	0.16	0.90	0.81	TC
38	0.09	0.281	2.972	2.898	2.79	0.27	0.94	0.45	SC
39	0.12	0.325	2.361	2.610	2.33	0.24	0.92	0.58	SC
40	0.19	0.394	1.580	0.502	1.60	0.20	1.01	0.77	SC

Notes. $n = 209$. LM = labeling a map; MCQ = multiple choices question; MNSQ = mean square; SA = short answers; SC = sentence completion; TC = table completion.

[a]Two options must be chosen.

[b]Items 14 and 15 had the same stem and the participants had to choose one option to receive the mark for each.

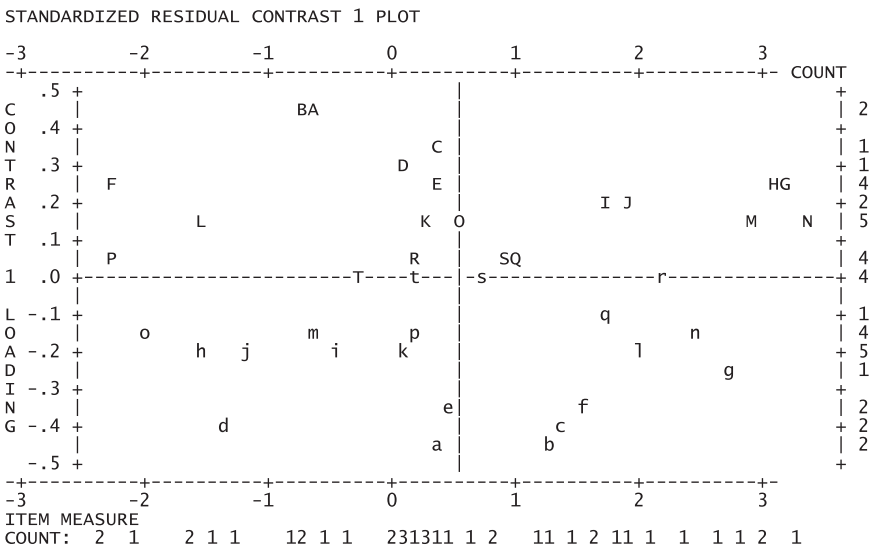

FIGURE 1 Illustration of the standardized residual contrast in the WLP Test. The horizontal line represents item difficulty measures and the vertical line represents contrast loading. Items are represented as alphabetic letters.

have not formed distinguishable patterns or clusters. Irrespective of their difficulty measure, they scatter in different regions of the map. This supports unidimensionality (Linacre, 2009b).

The first contrast in the residuals explained 3.4% (eigenvalue = 2.4) of the variance in the data. This is by far smaller than the 25.9% explained by the test items (eigenvalue = 18.4; the percentage expected by the Rasch model = 25.9%) and smaller than the 17.6% explained by person ability measures (eigenvalue = 12.4).

Finally, the eigenvalue of the first observed contrast in the simulated unidimensional data was 2.4 (rounded to 2), which is the same as the actual data set. Overall, this analysis supports the presumption that the data set is unidimensional—that is, targeted on the same underlying construct.

Rasch DIF Analysis

Previous Exposure to WLP Tests

I explored the items with significant UDIF for participants with and without previous experience of IELTS preparation courses. Items 8 and 9 were identified as cases of DIF. The difference between local difficulty magnitudes of the items is called the DIF contrast. For example, the local difficulty of test item 8 for test takers who had not taken preparation courses is -1.38 compared to -3.37 for test takers who had taken preparation courses. The difference between these two statistics exceeds 2.00, which is significant at $p < 0.01$ (Linacre, 2009b).

Items 8 and 9, which were the easiest test items (difficulty measure = -2.83), functioned in favor of the group with previous exposure to the test. Although their infit MNSQ values are between 0.5 and 1.5 (Linacre, 2009b), their outfit MNSQ values misfit (8 outfit MNSQ = 2.09; 9 outfit MNSQ = 2.68). Since outfit is sensitive to outliers, this indicates that some students with higher trait levels missed these easy items (Bond & Fox, 2007).

Next, both subgroups were segmented into high-ability and low-ability test takers, to investigate NUDIF with previous exposure to the IELTS exam as the grouping variable. WINSTEPS invoked 320 NUDIF comparisons (2 [lower ability groups] × 2 [higher ability groups] × 2 [grouping based on previous exposure] × 40 [items]), of which 41 (12.81%) had p values less than 0.05. In section 1 of the test, test items 3, 5, 6, 8, and 9 displayed 11 significant NUDIF instances (13.8% of the possible interactions in section 1 and 3.44% of the total of 320); in section 2, items 11, 17, 18, and 19 displayed nine significant NUDIF instances (11.25% of the possible interactions in section 2 and 2.81% of the total of 320); in section 3, item 21 displayed six significant NUDIF instances (7.5% of the possible interactions in section 3 and 1.88% of the total of 320); and in section 4, items 31, 32, 35, 37, and 40 displayed 15 significant NUDIF instances (18.75% of the possible interactions in section 4 and 4.69% of the total of 320).

MCQs 3, 5, 6, and 11 and table completion questions 8 and 9 functioned in favor of the low-ability segment of test takers with previous exposure to the IELTS listening test, several of whom were generally unlikely to answer MCQs accurately. Several low-ability test takers without previous exposure to the tests missed these items, with no common wrong answer identifiable in the responses. Items 31, 32, 35, 37, and 40 functioned in favor of high-ability test takers with previous exposure to the IELTS listening test.

Gender

I explored the UDIF analysis for gender. Two DIF items were identified: item 20 was easier for male participants, and item 23 was easier for female participants (both significant at $p < 0.01$). Standard errors of DIF were high for both classes in both items (joint SE = 1.11 and 1.03, respectively), and the contrasts were substantive (DIF contrasts of 1.62 and 2.43, respectively).

To investigate the NUDIF for gender, I defined two strata within each gender group: the upper ability and the lower ability. Again, 320 NUDIF analyses (2 [lower ability groups] × 2 [higher ability groups] × 2 [grouping based on previous exposure] × 40 [items]) were examined, of which 41 (12.81%) were statistically significant at $p < 0.01$ or $p < 0.05$. In section 1 of the test, test items 5, 6, 8, and 9 displayed six significant NUDIF instances (7.5% of the possible interactions in section 1 and 1.88% of the total of 320); in section 2, items 12, 15, 19, and 20 displayed 10 significant NUDIF instances (12.5% of the possible interactions in section 2 and 3.13% of the total of 320); in section 3, items 23 and 29 displayed eight significant NUDIF instances (10% of the possible interactions in section 3 and 2.50% of the total of 320); and in section 4, items 32, 33, 34, 35, and 37 displayed 17 significant NUDIF instances (21.25% of the possible interactions in section 4 and 5.31% of the total of 320).

Items 5, 6, 8, 9, and 20 (MCQ) functioned in favor of low-ability males; item 12, in favor of low-ability females; 15 and 23, in favor of high-ability females; and items 29, 32, 34, 35, and 37, in favor of high-ability males.

DIF Analysis for Nationality and Age

Interestingly, DIF analysis of nationality and age generated no significant DIF indices, suggesting that the operationalized construct did not function differently across different age and nationality groups.

DISCUSSION

This study set out to investigate DIF caused by gender, nationality, age, and degree of previous exposure to the test in a WLP test, using the logistic Rasch model and expert judgments. Although WLP tests have been reported to have undergone a rigorous process of development and revision (i.e., CEQPPC), the test users and researchers are left to presume that the tests are fair and test items do not favor any group of test takers. As previous studies suggest, it appears likely that this assumption is inaccurate. Overall, the results of this study suggest that item format and content of the WLP tests interact with certain characteristics of test takers and form bias in assessing their listening subskills. In this study, content analysis identified some test items as potential DIF items and unidimensionality analysis satisfied the preconditions for DIF analysis. In addition, in each test section, a number of test items were identified to interact with the aforementioned variables and caused DIF. The results are discussed in detail below.

Analysis of Test Content

Test items mainly focused on evaluating test takers' understanding of details and explicitly stated information, but a significant part of listening is the ability to make inferences, interpret illocutionary meaning, and draw conclusions (Buck, 2001; Shohamy & Inbar, 1991; Wagner, 2004). Geranpayeh and Taylor (2008) offer a similar description of the WLP tests developed by the University of Cambridge ESOL Examination Syndicate, especially IELTS. They describe the listening inputs as designed "with some internal repetition" and test items as "focusing on explicit and easily accessible information," with the intent of minimizing "any potential negative impact of hearing the text only once in slightly adverse conditions" (p. 3). Consequently, it appears that these tests represent the listening latent trait narrowly by focusing merely on the comprehension of details (Shohamy & Inbar, 1991). Shohamy and Inbar state that test items that compel test takers to concentrate on memory skills and insignificant details "tend to make severe demands on the test takers' memory load" and "should not be asked on LC [listening comprehension] tests" (p. 36).

The results of this study should be generalizable to WLP tests which are produced through the CEQPPC process, including IELTS. It seems that these tests do not measure "functional (pragmatic) knowledge," which test takers use in "understanding the function or illocutionary force of a statement or long text, and interpreting the intended meaning in terms of that" (Weir, 2005, p. 93). This function is higher order and requires interpreting and inference-making skills which go beyond literal meaning and paraphrasing (Hildyard & Olson, 1978). The probable construct-underrepresentation of WLP tests such as the IELTS listening test may help explain why test takers' scores on these tests and their later academic performance in the institutions of tertiary education have not been found to be significantly correlated (Ingram & Bayliss, 2007; Kerstjen &

Nery, 2000; Merrylees, 2003). Actual listening skills, particularly academic listening skills, are not as narrow as those measured in the tests, and scoring highly on the latter may not ensure or correlate with successful lecture, seminar, or tutorial comprehension (see Field, 2009).

These findings on test focus have implications and washback for language education (Alderson & Hamp-Lyons, 1996). Language instructors who help prepare students for WLP tests may restrict their instructional focus to developing students' ability to understand explicitly articulated information, techniques to deal with various item formats, and precursory skills to comprehension. Their choice of educational material is also affected by test structure (Bachman & Palmer, 1996). For example, Hayes and Read (2004) reported that IELTS preparation courses in a language school in New Zealand were "organized around the practice of skills, particularly through test-related tasks" (p. 109), although they did identify another school focusing on both the test and on interactional forms of language deemphasized by the test.

Second, institutions of tertiary education should be made aware of the limitations in the uses and interpretations of WLP test scores. While gaining a high score is desirable, it cannot in itself guarantee students' success in *listening-related* activities in universities, as studies on IELTS by Ingram and Bayliss (2007), Kerstjen and Nery, (2000), and Merrylees (2003) showed. The limited range of listening sub-skills examined in these tests should be considered when interpreting the test scores of applicants to institutes of higher education.

Dimensionality Analysis

Although content analysis identified some possible listening subskills evaluated by CEQPPC-based WLP tests, it did not predict how students would answer the test items or whether items would empirically tap the conceptualized dimension. I used the Rasch model to explore the dimensionality of items, item difficulty and test taker performance, and DIF. Dimensionality analysis indicated that responses are not contaminated by off-dimensional structures, as did different types of Rasch-based data analyses, such as infit analysis, PCAR, and a simulated study. However, some items have large outfit values, indicating that outlying (high- and low-ability) test takers did not perform as expected by the model. A DIF analysis was conducted to ascertain the degree to which there was an interaction between performance on these items and such variables as gender, age, nationality, and prior exposure to the test. This analysis showed that nationality and age had no significant impact on the testing behavior of participants, converging with IELTS claims about the internationality of the test. The effect of prior test exposure and gender are discussed below.

Effect of Previous Exposure to the Test

As noted above, several MCQs display DIF in favor of low-ability test takers with previous exposure to the IELTS listening test. The test's MCQs each have three options, so attempting a lucky guess has a high chance of success (33%). It appears likely that item distracters (the two incorrect answers) were attractive enough to distract low-ability test takers without prior exposure to the test, but those with prior exposure most likely made successful lucky guesses in answering the items. As Hayes and Read (2004) pointed out, test takers with previous exposure to a test are taught special test-taking strategies to answer questions, which might have helped them to answer items which their low-ability counterparts missed.

The lack of common wrong answers in the responses of the high-ability subgroup supports the supposition that they missed simple items out of carelessness. This assumption is supported by the outfit MNSQ patterns of this group: several incorrect answers on easy test items by high-ability had outfit MNSQ values greater than 1.5, indicating that their performance on these items was unexpected.

Some test items appear to have made listening-construct-irrelevant demands on test takers. For example, content analysis showed that takers could apply their comprehension of the oral input by paraphrasing the stem in item 5. If, however, test takers did not paraphrase the item stem, they would likely incorrectly answer the item even if they comprehended the audio input. This introduces a conceptually listening-irrelevant factor: Keeping the stem in mind while matching the audio stimuli to the written stimuli by paraphrasing seems to impose challenges beyond listening comprehension. Higher-ability students met these challenges well, as demonstrated by NUDIF analysis.

Items 17 and 18 functioned in favor of high-ability test takers with previous exposure to IELTS. Part of the oral input containing the answer to item 17 is ". . . and evening entertainment . . ." and test takers simultaneously read this short stem in their test booklets: "entertainment in the . . . [gap] . . . ." The item requires the rearrangement of the vocabulary in mind. Given the challenge of simultaneous reading of this particular item format, such rearrangement might impose more memory constraints on low-ability test takers, especially since the response to item 16 is located only a few words back in the oral message. This suggests that test item format is probably a constraint for low-ability people, which is consistent with Field's (2009) finding. This kind of item format demands the comprehension of too many details, which is a feature of many items in the test, and unfortunately precludes the deep comprehension of the test material (see Golen, 1990, for a discussion of listening barriers; see also Field, 2009).

The answer to item 18 is "(four-course) dinner," the phrase "four-course" being optional. Many low- and middle-ability test takers had written simply "four-course" or "full-course," which receives no partial credit, and several high- and low-ability test takers had written "dinner," which receives full credit. It may be that those who missed the items regarded writing all three words as the correct way of approaching the item (the instructions do not clearly say how many words should be written for this item but the word limit in all production items is three) but could not keep three words in mind, and that those who recalled the last word ("dinner") successfully put it down. This finding resonates with Bodie et al.'s (2011) study of the Watson–Barker Listening Test (WBLT), a commonly used test of listening comprehension. They discuss the implications of their study for constructing listening tests, as follows:

> Perhaps dichotomous scoring does not fully reflect listening ability, with the valid use of dichotomous scoring likely dependent on context. For instance, the section "understanding meaning" assumes that there is always only one correct meaning of a given utterance. Although there are certainly cases where this may be true . . . , in many interactions meaning can be as varied as the number of attendant listeners. Research across the academic landscape suggests that deriving meaning from conversation is more complex than picking out a single, correct meaning (for a review, see Edwards, 2011). Consequently, right–wrong scoring may misrepresent the multitude of meanings that may be viable alternatives. Indeed, a person who is able to generate multiple alternative meanings from a given utterance may be a more proficient or competent listener (see Burleson, 2011). (Bodie et al., 2011, p. 38)

This can be further explained in light of the logogen model of speech recognition. Logogens are abstract units storing a variety of information about a given word (e.g., semantic, orthographic, and phonemic properties) in the cognitive system. Each logogen becomes activated by the word whose attributes are consistent with the information stored in that logogen. The *orthographic* properties of a word, as stored in a logogen, can be loose and imprecise, despite the precision of other properties such as phonemic properties. For example, whereas a listener can recognize a word in its aural form by activating the related logogen (i.e., successful speech recognition), he or she might not be capable of spelling it accurately (i.e., failure in production) (see Morton, 1979). The IELTS listening test penalizes misspellings and ignores whether the test taker has recognized the word by activating the right logogen. The penalty for the misspelling can be reliable only when the rating benchmarks recognize and give weight to the activation of the right logogens, and this is achieved by using a polytomous scoring scale.

Further, item 18 apparently was meant to evaluate understanding details. That several test takers wrote "full" instead of "four" in their sheets showed that phoneme recognition skills, which contribute to but are different from listening comprehension, played a significant role. If this assumption is correct, item writers and test developers should note that "these narrow-spectrum, precursory-skill items may actually fail to tap the targeted comprehension construct directly," though they can have correlations with comprehension skills (Dunkel et al., 1993, p. 181).

Items 32, 35, 37, and 40 in section four of the test, which require written responses, are about Peregrine Falcons. For example, the answer to item 32 ("there is disagreement about their maximum . . .") is "(flight/flying) speed" and the oral input containing the answer is ". . . a number of textbooks claim that their flight speed can go as high as 350km an hour, so there is still some dispute about just how fast they can actually fly." This requires test takers to comprehend the oral input, paraphrase the input, apply their understanding to the item, and write the answer—three tasks taking place after comprehension.[1] It seems reasonable to assume that high-ability people with some test practice experience would outperform others on this series of tasks: Listeners in other subgroups might, in a freer condition, have understood that there was a dispute about the speed of these birds, but they failed to apply their understanding to the task. The same reasoning is applied to other items with significant DIF in this section.

[1] It has to be stressed that listening comprehension has been conceptualized from a variety of perspectives in "multidisciplinary fields" (Bodie, Worthington, Imhof, & Cooper, 2008). For example, Bodie et al. (2008) argued that listening can be viewed as: a) a function of personal attributes (e.g., willingness to listen and world knowledge), b) a function of cognitive processes (e.g., bottom-up/top-down processing), and c) a product (e.g., learning and establishing relationships). DIF analysis of listening tests focuses on the interaction of individual charactristics and stimuli features and informs our interepretation of the outcome of listening. A number of other perspectives to listening and its assessment emerge from the literature. For example, Bodie et al. (2011) investigated the constituent structure of the WBLT and reported that the test structure did not emerge as multi-factor, which contradicts the claims about the test structure in the literature. However, a discussion of the myriad conceptualizations of listening falls out of the scope of this article (for a review of different perspectives, see Beall, Gill-Rosier, Täte, & Matten, 2008; Bodie, 2009, 2010; Bodie & Fitch-Hauser, 2010; Bostrom, 2011; Buck, 2001; Floyd, 2010; Goh, 2005, 2010; Goh & Aryadoust, 2010; Imhof & Janusik, 2006; Janusik, 2007; Janusik & Wolvin, 2006; Liao, 2007; Lindahl, 2005; Schnapp, 2008; Smeltzer, 1993; Villaume & Bodie, 2007; Wolvin, 2010; Wolvin, Halone, & Coakley, 1999).

Effect of Gender

MCQs 5 and 6 and limited response items 8, 9, and 20 function in favor of low-ability males. These results are not surprising, given that 71% of male versus 49% of female test takers had taken IELTS preparation courses. Therefore, it seems likely that the gender factor interacted with previous exposure to the test, and that the findings about prior exposure can be generalized to the findings reported under gender. This finding resonates with Aryadoust et al.'s (2011) study of DIF in listening MCQs. They argued that in DIF cases favoring low-ability male participants:

> . . . a higher number of male than female low-ability test takers chose to randomly guess the correct response, a decision rewarded by the high success rate of guessing among three options. This analysis might suggest that these DIF items may not inherently function in favor of a subgroup but that a guessing factor related to gender is confounding the results. Cognitive psychologists note that male individuals in general tend to take more risks (in this case, attempting a lucky guess) when they encounter a problem . . . (Aryadoust et al., 2011, p. 18)

Aryadoust et al.'s (2011) suggested that test developers should safeguard the validity arguments of the test by, for example, increasing the choices in MCQs and writing attractive distracters.

However, item 12 functioned in favor of low-ability females, and items 15 and 23 functioned in favor of high-ability females. Reviewing these items again, we were convinced that the observed DIF in item 12 is most probably a statistical case, since the effect size is below 0.50. Item 15 is a MCQ written together with item 14; test takers are required to choose two out of five options. Although the answers are explicitly articulated in the oral input, it appears that the joint length of these two items and the options affected the lower ability students' performance. This observation resonates with Coleman and Heap's (1998) finding that asking two questions consecutively in the IELTS listening test may cause comprehension problems for the respondent. These items may measure test takers' reading speed and memory span, both irrelevant to the listening latent trait.

Limited production items 29, 32, 34, 35, and 37 functioned in favor of high-ability males. These items are highly difficult. Brindley and Slatyer (2002) reported that test takers "have about 30% less chance of being awarded the competency than if he or she were given [another item type than limited production]" (p. 380). Furthermore, some of these items are table completion items, which Coleman and Heap (1998) stated are the most difficult items for students in the IELTS listening test.

Limitations of the Study

This study is limited in scope because the test takers who attended the study were not drawn from a large international population. Additional research could focus on a larger population to augment the generalizability of the findings. Also, qualitative studies of test taking processes and strategies can shed light on the results of content and statistical DIF analysis; this study was limited in not employing that method. It is hoped that future research will further explore this least-researched area in assessing listening comprehension.

Although the administered test included table completion, MCQ, short answer items, sentence completion, and labeling item types, two item formats (i.e., classifying and matching test items) were not included in this test form. Therefore, we cannot draw conclusions about these item formats on the basis of the findings of this study.

CONCLUSION

The study has three implications for assessing listening in WLP tests. First, the listening construct in the test was found to be underrepresented, which is probably an important cause of the lack of significant correlation between test results and academic performance observed in previous studies. Second, short answer items, which feature prominently on the test, are likely to be biased in favor of higher-ability listener subgroups in listening comprehension because of these test takers' ability to apply swiftly what they have understood. These items are probably not effective tools to measure the construct because of the limitations they impose on listeners (see Hsiao-fang, 2004). Finally, it is likely that MCQs with few (in this test, three) options encourage lucky guesses. Low-ability people who have received training in test-taking strategies appear to be taking advantage of this fact, leading to flawed test results.

Because the test examined in this study was produced using the CEQPPC process, findings about it are likely to be generalizable to other high-stakes CEQPPC-developed WLP tests, and particularly to live versions the IELTS listening test. WLP test designers should be aware of the consequences and washback of designing a test exclusively around "items focusing on explicit and easily accessible information" (Geranpayeh & Taylor, 2008, p. 3), and test score users should consider the limitations of WLP test score uses and interpretations.

ACKNOWLEDGMENTS

I am grateful to the reviewers of *The International Journal of Listening* and Fred Meyer of the Fletcher School at Tufts University for providing valuable comments on this article; to my colleagues who helped me conduct the content analysis; and to the institutions that made the study possible. An earlier version of this study was presented at the CRPP's Conference (Redesigning Pedagogy: Designing New Learning Contexts for a Globalising World), 2009.

REFERENCES

Abbott, M. L. (2007). A confirmatory approach to differential item functioning on an ESL reading assessment. *Language Testing*, *24*, 7–36.

Alderson, C., & Hamp-Lyons, L. (1996). TOEFL preparation courses: A study of washback. *Language Testing*, *13*, 280–297.

American Educational Research Association/American Psychological Association/National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

Aryadoust, V. (2011a). Constructing validity arguments for the speaking and listening modules of international English language testing system. *The Asian ESP Journal*, *7*(2), 28–54.

Aryadoust, V. (2011b). Application of the fusion model to while-listening performance tests. *Shiken: JALT Testing & Evaluation SIG Newsletter*, *15*(2), 2–9.

Aryadoust, V., & Goh, C. (2010, December). Using a two-parameter logistic item response theory model to validate the IELTS listening test. *Proceedings of the Applied Linguistics Association of Korea (ALAK) International Conference, Seoul*, 210–215.

Aryadoust, V., Goh, C., & Lee, O. K. (2011). An investigation of differential item functioning in the MELAB Listening Test. *Language Assessment Quarterly*, *8*(4), 361–385.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford, England: Oxford University Press.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, England: Cambridge University Press.

Bachman, L. F., & Palmer, S. (1996). *Language testing in practice*. Oxford, England: Oxford University Press.

Badger, R., & Yan, O. (2009). The use of tactics and strategies by Chinese students in the listening component of IELTS. In L. Taylor (Ed.), *IELTS research reports* (Vol. 9). Canberra: IELTS Australia, Pty Ltd & British Council.

Banks, C. (1999). *An investigation into age bias in PET* (Cambridge ESOL Internal Research and Validation Report No. 22). Cambridge, England: University of Cambridge.

Beall, M. L., Gill-Rosier, J., Täte, J., & Matten, A. (2008). State of the context: Listening in education. *The Journal of International Listening*, *22*, 123–132.

Bejar, I., Douglas, D., Jamieson, J., Nissan, S., & Turner, J. (2000). *TOEFL® 2000 listening framework: A working paper*. (TOEFL Research Report No. RM-00-07, TOEFL-MS-19). Princeton, NJ: Educational Testing Service.

Bodie, G. D. (2009). Evaluating listening theory: Development and illustration of five criteria. *The International Journal of Listening*, *23*, 81–103.

Bodie, G. D. (2010). Treating listening ethically. *The International Journal of Listening*, *24*, 185–188.

Bodie, G. D., & Fitch-Hauser, M. (2010). Quantitative research in listening: Explication and overview. In A. D. Wolvin (Ed.), *Listening and human communication in the 21st century* (pp. 46–93). Oxford, England: Blackwell.

Bodie, G. D., Worthington, D., & Fitch-Hauser, M. (2011). A comparison of four measurement models for the Watson-Barker Listening Test (WBLT)-Form C. *Communication Research Reports*, *28*, 32–42.

Bodie, G. D., Worthington, D. L., Imhof, M., & Cooper, L. (2008). What would a unified field of listening look like? A proposal linking past perspectives and future endeavors. *International Journal of Listening*, *22*, 103–122.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. London, England: Lawrence Erlbaum Associates.

Bostrom, R. N. (2011). Rethinking conceptual approaches to the study of "listening." *The International Journal of Listening*, *25*, 10–26.

Breland, H., Lee, Y.-W., Najarian, M., & Muraki, E. (2004). *An analysis of the TOEFL CBT writing prompt difficulty and comparability of different gender groups* (ETS Research Rep. No. 76). Princeton, NJ: Educational Testing Service.

Brindley, G. (1998). Assessing listening abilities. *Annual Review of Applied Linguistics*, *18*, 171–191.

Brindley, G., & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing*, *19*, 369–394.

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

Burleson, B. R. (2011). A constructivist approach to listening. *The International Journal of Listening*, *25*, 27–46.

Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing*, *2*, 155–163.

Coleman, G., & Heap, S. (1998). *The misinterpretation of directions for the questions in the Academic Reading and Listening sub-tests of the IELTS test*. (Research Report No. 1, IELTS Australia).

Conoley, C. A. (2004). *Differential item functioning in the Peabody Picture Vocabulary Test - Third Edition: Partial correlation versus expert judgment*. Conoley, Colleen Adele: Texas A&M University.

Cotton, F., & Conrow, F. (1998). An investigation of the predictive validity of IELTS amongst a sample of international students studying at the University of Tasmania. In S. Wood (Ed.), *IELTS research reports* (Vol. 1; pp. 72–115). Canberra, Australia: IELTS Australia, Pty Ltd. & British Council.

Du, Y. (1995). When to adjust for differential item functioning. *Rasch Measurement Transactions*, *9*, 414.

Dunkel, P., Henning, G., & Chaudron, C. (1993). The assessment of a listening comprehension construct: A tentative model for test specification and development. *Modern Language Journal*, *77*, 180–191.

Edwards, R. (2011). Listening and message interpretation. *The International Journal of Listening*, *25*, 47–65.

Elder, E. (1996). The effect of language background on "foreign" language test performance: The case of Chinese, Italian, and modern Greek. *Language Learning*, *46*, 233–282.

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*, 113–148.

Field, J. (2009). A cognitive validation of the lecture-listening component of the IELTS listening paper. In L. Taylor (Ed.), *IELTS research reports* (Vol. 9). Canberra, Australia: IELTS Australia, Pty Ltd. & British Council.

Floyd, J. J. (2010). Listening: A dialogic perspective. In A. D. Wolvin (Ed.), *Listening and human communication in the 21st century* (pp. 128–140). Oxford, England: Blackwell.

Geranpayeh, A. (2001). *Country bias in FCE listening comprehension* (Cambridge ESOL Internal Research and Validation Report No. 271). Cambridge, England: University of Cambridge.

Geranpayeh, A., & Kunnan, A. J. (2007). Differential item functioning in terms of age in the certificate in advanced English examination. *Language Assessment Quarterly*, *4*, 190–222.

Geranpayeh, A., & Taylor, L. (2008). Examining listening: Developments and issues in assessing second language listening. *Cambridge Research Notes*, *32*, 3–5.

Goh, C. (2005). Second language listening expertise. In K. Johnson (Ed.), *Expertise in second language learning and teaching* (pp. 64–84). Basingstoke, England: Palgrave Macmillan.

Goh, C. (2010). Listening as process: Learning activities for self-appraisal and self-regulation. In N. Harwood (Ed.), *Materials in ELT: Theory and practice* (pp. 179–206). Cambridge, England: Cambridge University Press.

Goh, C., & Aryadoust, V. (2010). Investigating the construct validity of MELAB listening test through the Rasch analysis and correlated uniqueness modelling. In *Spaan fellowship working papers in second of foreign language assessment, 8* (pp. 31–68). Ann Arbor, MI: University of Michigan English Language Institute.

Golen, S. (1990). Factor analysis of barriers to effective listening. *Journal of Business Communication*, *27*(1), 25–36.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement, Issues and Practices*, *23*, 17–27.

Hansen, C., & Jensen, C. (1994). Evaluating lecture comprehension. In J. Flowerdew (Ed.), *Academic listening: Research perspective* (pp. 241–268). Cambridge, England: Cambridge University Press.

Hayes, B., & Read, J. (2004). IELTS test preparation in New Zealand: Preparing students for the IELTS academic module. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods* (pp. 97–111). Mahwah, NJ: Lawrence Erlbaum Associates.

Hildyard, A., & Olson, D. (1978). Memory and inference in the comprehension of oral and written discourse. *Discourse Processes*, *1*, 91–107.

Hsiao-fang, C. (2004). A comparison of multiple-choice and open-ended response formats for the assessment of listening proficiency in English. *Foreign Language Annals*, *37*, 544–555.

Imhof, M., & Janusik, L. A. (2006). Development and validation of the Imhof-Janusik listening concepts inventory to measure listening conceptualization differences between cultures. *Journal of Intercultural Communication Research*, *35*(2), 79–98.

Ingram, D., & Bayliss, A. (2007). IELTS as a predictor of academic language performance, part 1. In L. Taylor (Ed.), *IELTS research reports* (Vol. 7; pp. 137–204). Canberra, Australia: IELTS Australia, Pty Ltd. & British Council.

Jakeman, V., & McDowell, C. (2001). *IELTS practice test plus*. Harlow, England: Longman.

Janusik, L. (2007). Building listening theory: The validation of the conversational listening span. *Communication Studies*, *58*, 139–156.

Janusik, L. A., & Wolvin, A. D. (2006). *24 hours in a day. A listening update to the time studies*. Paper presented at the meeting of the International Listening Association, Salem, OR.

Kerstjen, M., & Nery, C. (2000). Predictive validity in the IELTS test. In R. Tulloh (Ed.), *IELTS research reports* (Vol. 6; pp. 85–108). Canberra, Australia: IELTS Australia, Pty Ltd.

Kim, Y. H., & Jang, E. E. (2009). Differential functioning of reading sub-skills on the OSSLT for L1 and ELL students: A multidimensionality model-based DBF/DIF approach. *Language Learning*, *59*, 825–865.

Kunnan, A. J. (1995). *Test taker characteristics and test performance: A structural modeling approach*. Cambridge, England: Cambridge University Press.

Liao, Y.-F. (2007). Investigating the construct validity of the grammar and vocabulary section and the listening section of the ECCE: Lexico-grammatical ability as a predictor of L2 listening ability. In *Spaan Fellow working papers in second or foreign language assessment, 5* (pp. 37–78). Ann Arbor, MI: University of Michigan English Language Institute.

Linacre, J. M. (1998). Structure in Rasch residuals: Why principal components analysis (PCA)? *Rasch Measurement Transactions*, *12*, 636.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*, 878.

Linacre, J. M. (2005). Rasch dichotomous model vs. one-parameter logistic model. *Rasch Measurement Transactions*, *19*, 1032.

Linacre, J. M. (2009a). WINSTEPS Rasch Measurement (Version 3.68) [Computer software]. Chicago, IL: Winsteps.com.

Linacre, J. M. (2009b). *A user's guide to WINSTEPS*. Chicago, IL: Winsteps.com.

Linacre, J. M., & Wright, B. D. (1994). Chi-Square fit statistics. *Rasch Measurement Transactions*, *8*, 350.

Lindahl, K. (2005). *How does God listen?* Woodstock, VT: Skylight Paths Publishing.

Merrylees, B. (2003). An impact study of two IELTS user groups: Candidates who sit the test for immigration purposes and candidates who sit the test for secondary education purposes. In R. Tulloh (Ed.), *IELTS research reports* (Vol. 4; pp. 1–58). Canberra, Australia: IELTS Australia Pty. Limited.

Morton, J. (1979). Facilitation in word recognition: Experiments causing change in the logogen model. In P. A. Kolers, M. E. Wrolstad, & H. Bouma (Eds.), *Processing visible language 1* (pp. 259–268). New York, NY: Plenum.

Muraki, E. (1999). Stepwise analysis of differential item functioning based on multiple-group partial credit model. *Journal of Educational Measurement*, *36*, 217–232.

Official IELTS Practice Materials. (2007). Available from www.IELTS.org

Powers, D. (1985). *A survey of* academic *demands related to listening skills*. (TOEFL Research Report No. 20). Princeton, NJ: Educational Testing Service.

Prieto, L., Alonso, J., & Lamarca, R. (2003). Classical test theory versus Rasch analysis for quality of life questionnaire reduction. *Health Quality Life Outcomes*, *28*, 1–27.

Raîche, G. (2005). Critical eigenvalue sizes in standardized residual principal components analysis. *Rasch Measurement Transactions*, *19*, 1012.

Richards, J. C. (1983). Listening comprehension: Approach, design, procedure. *TESOL Quarterly*, *17*, 219–239.

Schnapp, D. C. (2008). Listening in context: Religion and spirituality. *The International Journal of Listening*, *22*, 133–140.

Shohamy, E., & Inbar, O. (1991). Construct validation of listening comprehension tests: The effect of text and question type. *Language Testing*, *8*, 23–40.

Smeltzer, L. R. (1993). Emerging questions and research paradigms in business communication research. *The Journal of Business Communication*, *30*(2), 181–198.

Smith, R. M. (1996). Polytomous mean-square fit statistics. *Rasch Measurement Transactions*, *10*, 516–517.

Tennant, A., & Pallant, J. F. (2007). DIF matters: A practical approach to test if differential item functioning makes a difference. *Rasch Measurement Transactions*, *20*, 1082–1084.

The University of Cambridge ESOL Examination Syndicate Web site. (n.d.). IELTS teaching resources. Retrieved from http://www.cambridgeesol.org/teachingresources

The Web site of IELTS. (n.d.a.). Analysis of test data. Retrieved from http://ielts.org/researchers/analysis_of_test_data.aspx

The Web site of IELTS. (n.d.b.). Official IELTS practice materials order form. Retrieved from http://www.ielts.org/pdf/IELTS_SpecMatsOrderFormJan05.pdf

Tsui, A. B. M., & Fullilove, J. (1998). Bottom-up or top-down processing as a discriminator of L2 listening performance. *Applied Linguistics*, *19*, 432–451.

Villaume, W. A., & Bodie, G. D. (2007). Discovering the listener within us: The impact of traitlike personality variables and communicator style on preferences for listening style. *International Journal of Listening*, *21*, 102–123.

Wagner, A. (2002). Video listening tests: A pilot study. *Working Papers in TESOL and Applied Linguistics, Teachers College, Columbia University, 2*(1). Retrieved from http://journals.tclibrary.org/index.php/tesol/article/view/7/8

Wagner, A. (2004). A construct validation study of the extended listening sections of the ECRE and MELAB. *Spaan Fellow working papers in second or foreign language assessment*, *2*, 1–23.

Wagner, E. (2010). The effect of the use of video texts on ESL listening test-taker performance. *Language Testing*, *27*, 493–513.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago, Illinois: MESA Press.

Wolvin, A. D. (2010). Listening theory. In A. D. Wolvin (Ed.), *Listening and human communication: 21st century perspectives*. Oxford, England: Blackwell.

Wolvin, A. D., Halone, K. K., & Coakley, C. G. (1999). An assessment of the "intellectual discussion" on listening theory and research. *International Journal of Listening*, *13*, 111–129.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*, 370.

Weir, C. J. (2005). *Language testing and validation: An evidenced-based approach*. Basingstoke, England: Palgrave Macmillan.

APPENDIX
The IELTS Listening Subskills and Pertinent Items

Listening Subskills Evaluated by the WLP TESTS	*Definition and Description of the Subskills*	*Items Corresponding the Subskills*
Linguistic repertoire	The preacquainted knowledge of grammar, vocabulary, syntax, phonology, and morphology.	We agreed that all items rely on this skill. Yet, some items require more difficult processing of vocabulary heard and activating pertinent vocabulary units in brain. For example, the answer to item X is "pink slip." The juncture between the two words would be unclear to many test takers who are prone to miss the item; it appears that context to identify the juncture and the knowledge of phonology to tell apart "slip" from "sleep" are most important in this item.
World knowledge sources (schema)	Cultural and social knowledge organized in the form of mental structures in brain and contains information about social interactions.	We agreed that all items rely on this skill. If the listener does not realize how the state of the affairs and relationships operate, s/he is unlikely to be capable of processing and interpreting information.
Ability to make paraphrases	Items evaluating this subskill deal with comprehending vocabulary and structure and also need the listener to paraphrase the stimuli.	3, 4, 11, 13, 23, 31, 36, 37.
Ability to understand specific factual information such as names, numbers, and so forth	The ability to process the surface structure of the message mainly using linguistic knowledge. The listener is involved in detecting fillers or very specific pieces of information, such as names, dates, numbers, and so forth.	1, 2, 5, 6, 8, 9, 10, 12, 14, 15, 16, 17, 18, 19, 20, 21, 22, 24, 26, 27, 28, 29, 30, 32, 33, 34, 35, 39, 39, 40. The phrase containing the answer to Items 10 and 25 was repeated. The answer to item 9 was the age of a man in years, as opposed to other items which required detecting and writing specific words or choosing an option accordingly.
Integrate listening ability and visual skills	Items which have visual stimuli like a map or an object which test takers should name its parts.	1, 2. These items require that test takers look up a place on a map. They should understand directions and the description of dimensions.
Integrate listening, reading, short-term memory span, and/or writing abilities.	Listeners should understand the oral input, read the item, and put down the answer, which must not be longer than three words. Longer answers are not granted marks even though they contain correct information.	7, 8, 9, 10, 16, 20, 21, 22, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40.