Routledge
Taylor & Francis Group

International Journal of Listening

ISSN: 1090-4018 (Print) 1932-586X (Online) Journal homepage: https://www.tandfonline.com/loi/hijl20

What’s in a Topic? Exploring the Interaction Between Test-taker Age and Item Content in High-Stakes Testing

Jayanti Banerjee & Spiros Papageorgiou

To cite this article: Jayanti Banerjee & Spiros Papageorgiou (2016) What’s in a Topic? Exploring the Interaction Between Test-taker Age and Item Content in High-Stakes Testing, International Journal of Listening, 30:1-2, 8-24, DOI: 10.1080/10904018.2015.1056876

To link to this article: https://doi.org/10.1080/10904018.2015.1056876

The Intl. Journal of Listening, 30: 8–24, 2016
Copyright © International Listening Association
ISSN: 1090-4018 print / 1932-586X online
DOI: 10.1080/10904018.2015.1056876

What's in a Topic? Exploring the Interaction Between Test-taker Age and Item Content in High-Stakes Testing

Jayanti Banerjee
Worden Consulting LLC

Spiros Papageorgiou
Educational Testing Service

The research reported in this article investigates differential item functioning (DIF) in a listening comprehension test. The study explores the relationship between test-taker age and the items' language domains across multiple test forms. The data comprise test-taker responses (N = 2861) to a total of 133 unique items, 46 items of which were shared across two or more forms. Twenty-one items demonstrated DIF. However, there was no pattern by language domain. Eleven items showing DIF appeared in more than one test form but DIF for these items occurred only once.

Listening comprehension in a second language (L2) is a highly complex, individual, and interactive process (Vandergrift, 2007; Vandergrift & Tafaghodtari, 2010). During this process, L2 listeners use a variety of skills and strategies, including their ability to process the acoustic input, their language knowledge, and their world knowledge, to create an interpretation of the aural input (Buck, 2001; Vandergrift & Goh, 2012). As Buck (2001) argues "listening comprehension is not simply a process of decoding language. Meaning [. . .] is constructed by the listener in an active process of inferencing and hypothesis building" (p. 29). Interpretations of listening input can be expected to vary from listener to listener and, in an authentic listening context, these variations are unproblematic. If necessary, understanding can be checked and negotiated.

Second language (L2) listening assessments, however, look very closely at what test takers have understood from a stimulus, with little if any room for variations in interpretation. Additionally, large-scale and standardized listening tests are noncollaborative. The speaker is a disembodied voice from whom the test taker cannot seek a clarification or a reformulation. In this listening context non-linguistic knowledge such as background knowledge, past experience, feelings, and intentions can be a source of construct-irrelevant variance that could adversely affect test scores (Elliott & Wilson, 2013). For designers of L2 listening comprehension tests, examining such factors is critical because of their potential impact on test-taker performance (Brindley & Slatyer, 2002). If a test taker is unable to answer a listening comprehension question because

Correspondence concerning this article should be addressed to Jayanti Banerjee, Worden Consulting LLC, 115 Worden Avenue, Ann Arbor, MI 48103-4031. E-mail: j.v.banerjee@gmail.com

he lacks knowledge of the topic or the domain rather than because he cannot understand the language in the stimulus, then the item will fail to give appropriate information about that test taker's second language listening ability.

Indeed, a central concern in language test design and development is that test results (the scores assigned) must help us make appropriate decisions about test takers. Failure in this respect can have harmful consequences for individual test takers and for society. It is therefore the responsibility of test developers to ensure that test items operationalize the test construct and do not introduce construct-irrelevant bias. Item bias, in particular, is a threat to test fairness because test scores are affected by test-taker background characteristics such as gender and first language (L1) background rather than the construct that the test purports to measure (e.g., language ability). Differential item functioning (DIF) analysis is a common way of exploring test bias and of ruling out the possibility that items are systematically biased against particular groups of test takers. This type of analysis identifies item response patterns that can be attributed to group membership, i.e. cases where groups at the same level of ability have different probabilities of answering an item correctly (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014).

The study reported in this article explores the effect of L2 background knowledge on performance on a listening comprehension test of English as a second language. It is widely accepted that if the L2 listener has no background knowledge relevant to what is presented in an aural text, then it will be more difficult to make inferences based on such knowledge; thus, understanding of the input will rely primarily on linguistic knowledge, which can be challenging for many L2 listeners (Buck, 2001). The study employs a DIF approach to investigate the comprehension of items that test listening comprehension of spoken interactions in different L2 language domains by different age-groups. Our investigation focuses in particular on whether younger test takers might find the items from the occupational domain disproportionately challenging because they lack knowledge and understanding of the work world. This challenge might be exacerbated because of the nature of listening tests that are administered as part of standardized language exams: the stimuli are typically heard once and test takers are not in control of the pace at which they progress through the material. If DIF is observed for specific items that appear in more than one test form, the study investigates whether it is consistent across different test forms, the hypothesis being that items that consistently demonstrate DIF present a greater challenge to test construct and test fairness.

DIF in Language Tests: A Summary of Investigations to Date

DIF analysis is used to identify items that systematically function differently for test takers who share the same latent ability. It is important to note, however, that items demonstrating DIF are not automatically biased or unfair. Such items need to be carefully examined in case they demand knowledge and skills that are not relevant to the construct being assessed by the test. The items can be characterized as biased only if they tap construct-irrelevant abilities such that they could lead to unfair decisions being taken about the test takers. In order to examine DIF, it is first necessary to define the group of interest. This becomes the focal group (Zumbo, 2007). Typically the test population is split into two groups (e.g., male and female) but multiple groups are possible (e.g. three different age groups).

In second language assessment, there is a growing literature investigating DIF in relation to:

- language background such as speakers of Asian versus European languages (Kim, 2001); Indo-European versus non-Indo-European (Ryan & Bachman, 1992); Spanish versus Chinese (Chen & Henning, 1985; Sasaki, 1991); Arabic and Russian learners of Hebrew (Allalouf & Abramzon, 2008); and Mandarin versus Arabic (Abbott, 2007);
- academic background such as Humanities versus Sciences high-school seniors (Pae, 2004);
- gender (Aryadoust, Goh, & Kim, 2011; Pae & Park, 2006; Takala & Kaftandjieva, 2000); and
- accent familiarity (Harding, 2012).

The tests studied range from national tests such as the Finnish Foreign Language Certificate Examination (Takala & Kaftandjieva, 2000) to large-scale standardized tests such as the Michigan English Language Assessment Battery (Aryadoust et al., 2011) and the Certificate in Advanced English (Geranpayeh & Kunnan, 2007). They include tests of speaking (Kim, 2001), vocabulary (Takala & Kaftandjieva, 2000), listening (Harding, 2012), and reading (Allalouf & Abramzon, 2008). In each of these studies, the age group of the test takers was controlled to better isolate the variable of interest. For example, Pae (2004) studied high school seniors, Sasaki (1991) studied college students, and Takala and Kaftandjieva (2000) studied adults.

Investigations of DIF in Relation to Age for L2 Listening Comprehension Tests

When test takers of diverse ages take a test, examination providers need to ensure that the test takers' age does not affect their ability to respond correctly to the test items. Interestingly, a limited number of studies have investigated DIF in relation to the age of the test takers (Aryadoust, 2012; Geranpayeh & Kunnan, 2007). Geranpayeh and Kunnan (2007) explored age-related DIF in the context of a single test form of the listening section of the Certificate of Advanced English. They defined three age groups: the target age group for the exam (test takers in the 18–22 age range), a younger group (17 and younger), and an older group (23 and older). Geranpayeh and Kunnan (2007) used the Item Response Theory (IRT) computer program BILOG MG to analyze the items for DIF. They also asked content experts to analyze all the items and rate them for content bias. The content experts used a 5-point scale (ranging from strongly advantaged to strongly disadvantaged). They considered each item with respect to the three age-groups that had been defined and provided a judgment of whether the item was an advantage/disadvantage for each age group of test takers. The aim of this two-pronged approach was to isolate items that both demonstrated DIF (in the statistical analysis) and were thought to be an advantage/disadvantage for a particular group of test takers. Such items could be considered biased.

The statistical analysis revealed that six out of the 32 items administered exhibited DIF in relation to the target age (reference) group. Of these six items, one item exhibited DIF for both the focal groups (i.e., the group that was 17 and younger, and the group that was 23 and older); two exhibited DIF for the younger age group; and the remaining three exhibited DIF for group 3 (23 and older). The content analysis identified seven items that could disadvantage younger test takers and no items that could disadvantage the older test takers. However, the experts' judgments only corresponded to the DIF analysis in the case of one item. This is perhaps unsurprising; Alderson (1990) and more recently Alderson and Kremmel (2013) found that content experts (experienced English-as-a-Second-Language teachers) were unable to agree with one another on what an item is testing. It stands to reason, therefore, that expert judgments of content bias might not correspond to the results of DIF analysis. That said, the meaning of DIF statistics is not self-evident and it is unwise to rely only on either statistical analyses or on human judgments. The two must be contrasted and compared.

Looking more closely at their results, Geranpayeh and Kunnan (2007) hypothesized that DIF could be attributed to test takers' cognitive process, particularly in their ability to recall information and their ability to use memory strategies. They also suggested that another possible reason for the presence of DIF was that the items were multidimensional, i.e. they measured a different part of the relevant construct in relation to the other items on the test. If the latter hypothesis could be confirmed, it would suggest that there is no construct-irrelevant bias (cf. Pae, 2004) and that, far from introducing a challenge that is potentially unfair to a group of test takers, the test is (rightly) tapping a richer construct. Ultimately, Geranpayeh and Kunnan (2007) were unable to fully confirm or deny age-related bias.

In his study of a practice test form of the listening section of the International English Language Testing System (IELTS), Aryadoust (2012) investigated DIF with respect to test taker age, along with gender, nationality, and previous exposure to the test. Building on previous work (Aryadoust et al., 2011), Aryadoust was interested in the interaction between test-taker ability and their gender, nationality, and/or previous exposure to the IELTS. Consequently, he explored both uniform DIF (whereby for two sub-groups of test takers the amount of DIF is constant across ability levels) and nonuniform DIF (whereby the DIF varies across ability levels). Aryadoust (2012) found a substantial number of interactions between gender and ability level (NUDIF): five items favored low-ability male test-takers; five items favored high-ability male test takers; one item favored low-ability female test takers; and two items favored high-ability female test-takers. Aryadoust (2012) was unable to pinpoint the features of the items that caused them to demonstrate DIF. Furthermore, he found no items demonstrating nationality- or age-related DIF.

STUDY RATIONALE

Despite these studies, there is still scope for further investigation of DIF in relation to age groups taking a listening test in order to better understand the effect of background knowledge on performance on tests of L2 listening comprehension. One area that warrants further investigation is the relationship between the age of the test takers and the context of the language activities, that is, whether the language interaction would be expected to occur in the personal, social, educational, or occupational domains (Council of Europe, 2001). Although care should be taken so that test content is accessible to all test takers regardless of their world experience, test takers without direct work experience might find the items from the occupational domain disproportionately challenging because they lack knowledge and understanding of the work world. Therefore, items in the occupational domain might be more prone to demonstrating DIF with test takers who are adolescents and young adults. It could also be further assumed that this is more likely to happen when there is more domain-specific language in the listening input, for example test takers are asked to comprehend a conversation between two colleagues discussing the completion of a timesheet, as opposed to more general language, for example when two colleagues discuss transportation to work. Such a study would be particularly unique because it proposes confirmatory DIF analyses which test a hypothesis about the direction of the DIF (see Ferne & Rupp, 2007). Indeed, studies to date have primarily (though not exclusively) taken an exploratory approach; first items demonstrating DIF are identified and then these items are analyzed for patterns. This study aims to contribute to the relevant literature by adopting a confirmatory DIF approach. One advantage of this theory-driven approach is that its starting point is a hypothesis about the data. In our case this hypothesis is that test items from a specific language domain (in this case occupational) will demonstrate DIF with younger test takers.

An additional gap in research to date is the exploration of DIF for items that are reused in subsequent test administrations. We acknowledge Linacre (2011) who points out that DIF demonstrated by one item, even when it is large, will have a negligible effect on person ability estimates using a Rasch model analysis. We also accept that the effect of DIF for individual items might cancel out at the test level (Pae & Park, 2006). However, if an item were to systematically demonstrate DIF on every occasion that it is included in an exam, this is a much more serious concern because an advantage or disadvantage for a specific group is perpetuated. Therefore, investigation of DIF across multiple test forms is essential.

With this in mind, the main objective of this study was to investigate the effect of test-taker age upon listening test performance in a multi-level English language proficiency test with a particular focus on the item context/domain. The questions it addressed were:

1. Do items from a specific language domain (in this case occupational) demonstrate DIF with younger test takers?
2. If DIF is observed, is it consistent across different test forms?

METHOD

The Michigan English Test

The Michigan English Test (MET) is a multilevel test of general English language proficiency intended for adults and adolescents at or above a secondary level of education and is typically used for educational purposes, such as when finishing an English language course, or for employment purposes, such as applying for a job or pursuing a promotion that requires an English language qualification (see http://www.cambridgemichigan.org/met for more information). The MET was chosen for this study for two main reasons. First, it is intentionally designed for a wide test-taker base. Second, the stimuli target all of the afore-mentioned language domains approximately equally. Data were collected from four unique test forms (test versions) during four operational administrations in 2009 and 2010. Reliability indicated by Cronbach's alpha ranged from .90 to .93 (Cambridge Michigan Language Assessments, 2014). Rasch reliability statistics are presented later in this article.

The listening section was of particular interest for this study. This contains 60 multiple-choice items, divided into three parts: short dialogues (D) between two interlocutors followed by one question, longer dialogues (S) with two interlocutors preceding three to four questions, and monologues (M) followed by four to five questions. Items have four answer options, with one key and three distracters. Questions and options are printed in the test booklets, and test takers are given the option to take notes when listening to the stimuli.[1] However, unlike an increasing number of tests on the market, the listening stimuli are played only once. Therefore, despite the option to take notes, the test is more characteristic of a while-listening performance test where test takers need to simultaneously read and answer the test items while listening to the stimuli (Aryadoust, 2012). This could exacerbate the interaction of a domain, or listening subskill with age because test takers cannot linger over the item (as would be the case for grammar and/or reading items). Nor do they get a second opportunity to listen to the input.

Participants

The test forms were administered to a total of 2861 test takers ($N = 672$, 618, 743 and 828, respectively) at nine exam centers in a Latin American country. A single country was selected in an attempt to control for a potential effect of the social and educational context. All test takers indicated that Spanish was their first language. This allowed us to control for the potential effect of language background.

Because this study focused on the effect of age upon performance in specific item types (occupational domain items), it was important to have unimpeachable definitions of age. In particular, it was important to delimit two groups, one of which was highly unlikely to have work experience (group 1) and the other of which was highly likely to have work experience (group 3). A middle group was defined in order to account for test takers who fell between the two categories and whose performance profiles might not be as revealing. The resulting definitions were:

Group 1: test takers below the legal school-leaving age, who could be expected to have little or no workplace knowledge

Group 2: test takers above the legal school-leaving age but who might not have entered the workforce, perhaps because they are studying at university. This group may have some workplace knowledge by virtue of holiday jobs or part-time jobs.

Group 3: test takers who had definitely entered the work-force, who could be expected to have substantial workplace knowledge.

The test centers from which the data had been collected provided the age ranges corresponding to these definitions. Group 1 included all test takers who were younger than 17 years old; group 2 included test takers 17–27 years; and group 3 included all test takers who were older than 27. Table 1 presents the distribution of the test takers by age group across all four forms:

The study population spanned the English proficiency levels A2—C1 as described in the CEFR.[2] Table 2 presents the distribution of test takers by age group and proficiency level.

The tables show that the majority of test takers are in Group 2; that is, they are older than 17 but younger than 27. The majority are also at the B1 level on the CEFR. As evidenced by the MET 2013 report (Cambridge Michigan Language Assessments, 2013), this distribution is representative of the general MET population.

[1] A sample test is available at http://www.cambridgemichigan.org/test-takers/prepare-study/.

[2] For more information on the relationship between MET test scores and the CEFR levels please refer to Papageorgiou (2010).

TABLE 1
Age ranges defined across all test forms

Group	*Age Range*	*Form A*	*Form B*	*Form C*	*Form D*	*Total (N)*
1	< 17	22.1%	11.5%	24.0%	7.6%	460
2	≥ 17 but ≤ 27	63.2%	64.1%	57.7%	61.1%	1756
3	> 27	14.7%	24.4%	18.3%	31.3%	645
Total (N)		**672**	**618**	**743**	**828**	**2861**

TABLE 2
Range of proficiency levels by age group

Group	*Age Range*	*Level A2*	*Level B1*	*Level B2*	*Level C1*	*Total (N)*
1	< 17	16.6%	18.3%	13.3%	10.6%	460
2	≥ 17 but ≤ 27	62.3%	60.9%	62.0%	59.9%	1756
3	> 27	21.1%	20.8%	24.7%	29.5%	645
Total (N)		**745**	**1245**	**579**	**292**	**2861**

Data

As it was pertinent to the aims of this study, the administrations selected contained repeated items, which serve test equating requirements as part of the routine administration of the test. A total of 184 items (responses to 46 items in each test form) were analyzed in this study; 43 items were repeated in more than one form (35 items used twice and 8 items used 3 times); 19 of these belonged to the occupational domain. A total of 133 unique items were analyzed, that is, 90 items that occurred once and 43 items that appeared twice or three times. To address the research questions, these 133 items were classified by the domain of the language activity (personal, social, educational, and occupational). Initial classifications were available from the test development process, during which item writers routinely assign the targeted primary domain. To confirm these classifications, we independently re-classified each item. We then resolved discrepancies both with each other and with the original classifications until consensus on item classification was reached. The data comprised 42 occupational items, the focus of our analysis, and 91 nonoccupational items. Appendix A presents examples of one item type each (the short conversation) in the occupational, educational, and public domains. For examples of the long conversation and extended monologue item types, please refer to the sample MET test available on the CaMLA website (http://www.cambridgemichigan.org/test-takers/prepare-study/).

DIF Analysis

As we explained earlier, there are a number of alternative models available for exploring DIF. However, the Rasch model (Rasch, 1980) is frequently used (Aryadoust et al., 2011) and is particularly appropriate when datasets are small, as is the case in this study (for a discussion of appropriate N sizes of different IRT models, see McNamara, 1996, p. 295). Although there are some differences such as the estimation of item fit (Linacre, 2005), the Rasch model is similar to the one-parameter IRT model when analyzing dichotomous data. The Rasch model produces linear measures of item difficulty and person ability on a common interval scale of "log odds" units (McNamara, 1996, p. 165) centered on 0, the logit scale. Positive values indicate more proficient test takers or more difficult items, while negative values indicate less proficient test takers or easier items. Through analysis of the differences between observed responses and responses expected by the Rasch model, fit statistics are calculated, indicating the degree to which items fit the underlying construct, which is essential for conducting a Rasch-based DIF analysis (Aryadoust, 2012).

For the purposes of this study, the infit mean square statistic (Infit MNSQ) was inspected because of its reliance on responses of test takers whose ability is well-matched with item difficulty on the logit scale (Bond & Fox, 2007). The infit mean square statistic has an expected value of 1. Although interpretation of acceptable values depends on the data (Lincare & Wright, 1994), typically values above 1.3 show significant underfit. This indicates a lack of predictability, signaling either that the items are problematic or that they do not measure the same trait as other items in the test (Bond & Fox, 2007). Values below .75 show significant overfit, indicating a lack of variation. This suggests the overall response pattern is too predictable and there might be content overlap with other items (McNamara, 1996). As can be seen in Table 3, the values of the infit mean square statistic were acceptable: no items demonstrated significant underfit and only two items in Form B (with infit values .73 and .74) were very close to the lower threshold. Therefore, the measurement properties of the items were appropriate for the test taking population. Furthermore, the lack of items with significant underfit confirmed the psychometric unidimensionality of the data and their fit to the Rasch model. Table 3 also presents Rasch person and item reliability for each test form. Person reliability indicates how well the test discriminates the population into different levels of ability and is interpreted similarly to Cronbach's alpha (Linacre, 2011). Item reliability indicates the precision of the item measures on the underlying latent variable assessed by the test (Beglar, 2010).

Four separate analyses were run with the Rasch computer program WINSTEPS (Linacre, 2011) to examine whether items that appeared in more than one form would demonstrate DIF consistently across these forms (research question #2). This focus on individual items that were repeated across test forms precluded differential bundle functioning (DBF) analysis (Abbott, 2007).

WINSTEPS uses a logit-difference (logistic regression) method (Linacre, 2014) to present DIF results by estimating the difference between the Rasch item difficulties across pairs of groups

TABLE 3
Range of infit mean square statistics

Form	*Minimum Infit Mean Square*	*Maximum Infit Mean Square*	*Person Reliability*	*Item Reliability*
A	.77	1.26	.89	.98
B	.73	1.24	.91	.98
C	.82	1.30	.90	.98
D	.82	1.27	.90	.98

(e.g., Group 1 vs. Group 2, Group 1 vs. Group 3), matched by group ability (indicated by their performance on the whole test (Linacre, 2011)). Because this study focuses on the comparison between groups that are likely to differ in their workplace experience (see research question #1), the presentation of results in the next section differs from that of Geranpayeh and Kunnan's (2007), who treated the middle group (ages 18–22) as the reference group and examined DIF by comparing item difficulty for the youngest and the oldest groups separately in relation to the reference group. The DIF results of our study, as we show in the next section, compare the performance of Group 1 (youngest) and Group 3 (oldest) which are more likely to differ in their workplace experience. We do not treat one group as the reference group and the other group as the focal group in our analysis. Such analysis is typical of what Zumbo (2007, p. 224) calls "first generation" DIF studies, which labeled groups as the reference and focal ones to denote minority and majority groups respectively, typically based on gender or race. In our case we do not consider any group as the majority or minority group. Given our use of IRT and a confirmatory (theory-driven) approach, our study has elements of what Zumbo describes as "second generation" DIF studies. However, in terms of its conceptualization, our study can be seen as belonging to Zumbo's (2007) "third generation" DIF studies, because we conceive of DIF as possibly occurring because of some characteristic of the test item (in our case the occupational domain) that is not relevant to the underlying ability we want to access, in our case the listening comprehension of the younger test takers who took the MET.

As will be explained in the next section, DIF was investigated in relation to two measures for Groups 1 and 3:

- DIF contrast, which indicates the difference in logits units of the item difficulty when the item is taken by Group 1 and Group 3. Following Linacre (2011) we consider differences between 0.40 and 0.60 logits "slight to moderate" DIF contrast, and higher than 0.60 logits "moderate to large" DIF contrast.
- Probability, which indicates whether the observed DIF contrast is statistically significant. WINSTEPS provides a Welch t statistic, which is a two sided t- test for the difference between two DIF estimates (Group1 and Group 3), testing the null hypothesis that two DIF estimates are the same, except for the measurement error. A probability (p) value indicates whether the difference between the two DIF estimates is statistically significant (Aryadoust et al., 2011).

RESULTS

Table 4 summarizes the differences in ability between groups. For each test form the table shows, first for Group 1 and then for Group 3, the number of test takers (N), mean person ability measure expressed in logits and the standard error of the mean logit value. The mean person ability measure is estimated without the responses of test takers who responded correctly to all items or did not respond to any items. This exclusion is because, mathematically, the ability of persons with extreme (all items correct/wrong) scores is not directly estimable. The difference of the mean logit value (Group 1 person ability measure minus Group 3 person ability measure) and its standard error are also presented. The last three columns present information related to the Welch

TABLE 4
Differences in ability between Group 1 and Group 3

	Group 1			*Group 3*			*Difference*		*Welch*		
Form	*N*	*Mean logit*	*SE*	*N*	*Mean logit*	*SE*	*Logit*	*SE*	*t*	*df*	*Prob.*
A	148	−0.03	0.08	97	0.76	0.14	−0.80	0.16	−4.99	150	0.000
B	71	0.58	0.14	150	0.21	0.11	0.37	0.18	2.05	151	0.042
C	177	0.06	0.08	132	0.28	0.10	−0.21	0.13	−1.65	256	0.100
D	63	0.09	0.16	259	0.37	0.08	−0.28	0.18	−1.54	92	0.128

TABLE 5
DIF statistics for items that were more difficult for Group 1

				Group 1			*Group 3*			*DIF*		*Welch*		
Form	*Item ID*	*Domain*	*Other form*	*FV*	*DIF*	*SE*	*FV*	*DIF*	*SE*	*Contrast*	*SE*	*t*	*df*	*Prob*
A	D15	Occup	No	0.27	1.11	0.20	0.55	0.41	0.24	0.69	0.31	2.22	222	0.027
	S22	Public	No	0.34	0.73	0.19	0.63	−0.04	0.24	0.77	0.31	2.51	218	0.013
B	M43	Occup	Yes	0.39	1.12	0.28	0.55	−0.16	0.19	1.28	0.34	3.79	162	0.000
C	D6	Occup	No	0.47	0.18	0.17	0.65	−0.57	0.20	0.75	0.26	2.87	288	0.004
	S24	Public	No	0.42	0.41	0.17	0.58	−0.17	0.20	0.58	0.26	2.22	291	0.027
	S29	Educ	No	0.34	0.79	0.17	0.52	0.10	0.20	0.68	0.26	2.59	294	0.010
	M40	Occup	Yes	0.46	0.21	0.17	0.60	−0.32	0.20	0.53	0.26	2.06	288	0.041
D	D10	Educ	No	0.48	0.12	0.29	0.75	−1.05	0.16	1.16	0.33	3.53	138	0.001
	S29	Educ	No	0.46	0.20	0.29	0.66	−0.57	0.15	0.77	0.33	2.36	133	0.020

test, namely the value of the *t* statistic, the associated degrees of freedom (df) and the probability of observing the *t* statistic.

Table 4 shows that the person ability measure of Group 3 was higher on Forms A, C and D. Group 1 was more able on Form B. Although Group 3 in general appeared more able than Group 1, which gives support to the group ability matching performed by WINSTEPS for the DIF analysis of this study, the differences in ability were statistically significant ($p < .05$) only with Forms A and B.

Tables 5 and 6 present the items that demonstrated DIF; Table 5 lists the items that were found to be more difficult for Group 1 and Table 6 lists the items that were found to be more difficult for Group 3.

Both tables present information similar to that presented by Aryadoust et al. (2011):

- The first four columns provide the test form, the item type and position,[3] the domain tapped by the item, and, whether the item appeared in another form.

[3] The item ID indicates the type of item (D = short dialogues, S = longer dialogues, M = monologues) and the position of the item in the test booklet. Unless otherwise indicated, the same ID across different test forms only specifies the same item type and location in the booklet, not identical content.

TABLE 6
DIF statistics for items that were more difficult for Group 3

			Other	*Group 1*			*Group 3*			*DIF*		*Welch*		
Form	*Item ID*	*Domain*	*form*	*FV*	*DIF*	*SE*	*FV*	*DIF*	*SE*	*Contrast*	*SE*	*t*	*df*	*Prob*
A	S21	Public	Yes	0.44	0.24	0.18	0.43	1.05	0.24	−0.81	0.30	−2.66	213	0.008
	S20	Public	Yes	0.51	−0.14	0.18	0.51	0.64	0.24	−0.78	0.30	−2.60	214	0.010
B	D2	Occup	Yes	0.84	−1.49	0.35	0.62	−0.54	0.19	−0.96	0.40	−2.39	144	0.018
	D8	Educ	Yes	0.67	−1.21	0.32	0.55	−0.14	0.19	−1.07	0.37	−2.85	151	0.005
	D14	Educ	Yes	0.48	0.66	0.27	0.28	1.43	0.21	−0.77	0.35	−2.22	173	0.028
C	D10	Personal	Yes	0.55	−0.25	0.16	0.47	0.38	0.20	−0.63	0.26	−2.43	289	0.016
	S22	Public	No	0.75	−1.24	0.18	0.65	−0.57	0.20	−0.67	0.27	−2.46	295	0.015
	M38	Educ	No	0.58	−0.36	0.17	0.39	0.80	0.21	−1.15	0.26	−4.38	286	0.000
D	D15	Personal	No	0.49	0.03	0.29	0.34	1.21	0.15	−1.18	0.33	−3.62	135	0.000
	S26	Public	Yes	0.66	−0.30	0.29	0.46	0.44	0.14	−0.73	0.32	−2.29	131	0.024
	S32	Occup	Yes	0.56	−0.30	0.29	0.40	0.88	0.15	−1.17	0.32	−3.64	133	0.000
	S34	Occup	Yes	0.46	0.20	0.29	0.37	1.05	0.15	−0.85	0.33	−2.60	134	0.010

- The next three columns present information about the performance of Group 1, including the facility value (FV) of the item, i.e. the percent of test takers who responded correctly, the DIF measure (i.e., the Rasch item difficulty for Group 1 expressed in logits) and the standard error (SE) associated with the DIF measure.
- The same information for Group 3 is presented in the subsequent three columns.
- The DIF contrast in the eleventh column shows the DIF difference between the two groups. It is positive in Table 5 because items are more difficult for Group 1 but negative in Table 6 because items are more difficult for Group 3. The standard error (SE) of the DIF contrast is also shown.
- The last three columns present information related to the Welch test, which examines the probability of observing the amount of DIF contrast (Column 11) by chance, when there is no systematic item bias effect. The value of the t statistic, the associated degrees of freedom (df) and the probability of observing the t statistic are provided.

Both tables include only items with a statistically significant probability ($p < .05$) and following Linacre (2011, p. 361) "slight to moderate" (between .40 and .60) and "moderate to large" (higher than 0.60) DIF contrast.

The tables reveal a rather muddy outcome. Twenty-one items demonstrated DIF; 9 were more difficult for the youngest test takers (group 1) and 12 were more difficult for the oldest test takers (group 3). Out of the nine items that were more difficult for the youngest test takers (Table 5), four were set in the occupational domain, two in the public domain and three in the educational domain. Out of the 12 items that were more difficult for the oldest test takers (Table 6) 3 were set in the occupational domain, 4 in the public domain, 2 in the personal domain, and 3 in the educational domain. A chi-square test confirmed that the association between the group (younger test takers or older test takers) and domain (occupational items and non-occupational items) was not significant χ^2 (1) = .875, p = .350.

Eleven items demonstrated DIF in one test administration, but no DIF was present when the same items were included in one of the other three forms. The lack of recurrent statistically significant DIF for the same item across multiple test forms might have been due to the relatively small number of some groups of test takers, thus reducing the power and consistency of the DIF test statistic. If DIF provided information that should be ignored due to inconsistency across test forms, there should be no relationship between DIF results for the same item across DIF analyses of two test forms. A sign test[4] proved otherwise, as 73% ($p < .001$) of the time Group 1 found an item easier or more difficult than Group 3 across multiple occurrences of that item. Therefore, the answer to our second research question is inconclusive. While there was only a single item that had statistically significant DIF more than once, the direction of the DIF statistics showed a positive relationship between DIF results from one test form to another test form. So the inconsistency in identifying DIF may be due to the fact that there was no DIF in the first place and that the observation of DIF was purely by chance, in particular because of the relatively small sample size for some of the test taker groups.

DISCUSSION AND CONCLUSIONS

Regarding research question number 1 (Do items from a specific language domain (in this case occupational) demonstrate DIF with younger test takers?) the statistical analysis did not reveal a consistent pattern. Of the nine items that were more difficult for the youngest group, only four belonged to the occupational domain, whereas the remaining five did not. Moreover, 3 of the 12 items that were more difficult for the oldest group also belonged to the occupational domain. Therefore, our initial hypothesis was not confirmed; items in the occupational domain did not consistently demonstrate DIF with younger test takers. One explanation for this finding could lie in the item specifications for the MET. Items are intentionally crafted to ensure that specialized background knowledge is not required to respond to test items. Therefore, items are set in a “light” occupational context, with mostly general language being tested. This might result in items in the occupational domain being perfectly accessible to younger test takers.

Regarding research question number #2 (If DIF is observed, is it consistent across different test forms?) the analysis revealed “inconsistent” DIF. That is, although 11 out of the 21 items showing DIF appeared in more than one test form, DIF only occurred on one instance of item use. To explain this absence of DIF for the same item across test forms, we turn to Pae (2004) and Geranpayeh and Kunnan (2007). Pae (2004) suggests that a large DIF value indicates that an item measures an additional construct. Geranpayeh and Kunnan (2007) hypothesize that the item measures an auxiliary dimension differently across two test taker groups. This suggests that items demonstrated DIF in one form because, in that instance, they measured an auxiliary dimension that the other items in that form did not. However, this explanation of inconsistent DIF should be treated with caution because, as we have discussed in the results section, the inconsistency may be due to our small samples sizes. Additional work with larger test taker numbers will be needed to address the second research question of our study.

[4] For a description of the sign test see Abdi (2007).

The findings of this study have some important implications for developers of general proficiency tests of listening comprehension. As Harding (2012) points out, the interpretation of DIF as evidence of bias will depend on the purpose of the test and the nature of the target language use (TLU) domain. If it is important to understand English in a variety of contextual domains, the findings of this study suggest that if the tested language is not too domain-specific, then a variety of domains can (and, indeed, should) be included in the test without disadvantaging those groups that do not have experience in these domains. Additionally, even though younger learners might not have direct experience with workplace contexts, it is likely that they have indirect experience, e.g. when they visit their parents at work or when they watch movies and documentaries. This indirect experience might be sufficient to ensure that they are not disadvantaged when they answer listening comprehension items testing general language proficiency in a workplace context. It is also important to note that, in the Rasch model, DIF from a small number of items will have a negligible impact on estimates of person ability, especially with longer tests such as the listening section of the MET (see Linacre, 2011, p. 448).

It is also important to engage with the extent to which an item demonstrates consistent DIF across test forms. The small sample size in our study hampered our efforts to test this hypothesis but the large question remains. How might form assembly account for the fact that the same item might demonstrate DIF in one test form but not in another? We know that the effect of DIF for individual items might cancel out at the test level (Pae & Park, 2006). We also accept that the opposite effect is also possible, that is, that items stored in an item bank might be combined in such a way that they will result in an overall DIF effect at the test level (Takala & Kaftandjieva, 2000). However, if DIF is not persistent (i.e., occurs on only one instance of item use) then we cannot agree with Takala and Kaftandjieva's (2000) recommendation that items be retired as soon as DIF is detected. Additionally, we need to consider Geranpayeh and Kunnan's (2007) view that the combination of an item with other items might result in DIF that is indicative of an auxiliary dimension rather than bias. Taking all this into consideration, a closer examination of the content of items showing DIF is required during test form assembly. This is particularly the case if test developers, in an effort to ensure parallel test forms across time, wish to see the same auxiliary dimensions measured in each test form.

To better understand the sources of DIF in listening tests, it is also worth exploring the effect of the subskill tapped by items (i.e., global, local, or inferential). DIF might be observed for different test taker groups depending on the cognitive process that subskill demands. As Geranpayeh and Kunnan (2007) note, DIF could be attributed to test takers' cognitive process, particularly their ability to recall information and their ability to use memory strategies. It is conceivable that tasks which are more cognitively challenging will demonstrate DIF among younger test takers.

Naturally, the findings of this article should be interpreted with caution due to some limitations. The data were collected from administrations of one test (the MET), containing scripted stimuli that were only played once and with just one response type (multiple-choice options). The first language was also purposely constrained. Therefore the findings of this study might not generalize to listening tests that contain different item types or play the stimuli twice, or to other L1 groups. It should also be noted that N sizes for some test taker groups were in general small and unequal across age groups, which inevitably affected the precision of the person ability and item difficulty estimates of the Rasch model analysis and consequently estimates of DIF, in particular those estimates for repeated items (for which we claim that DIF was not consistent).

These are issues that warrant further research. Perhaps more important is the acknowledgement that the DIF statistics are only indicative of a possible problem but do not provide a transparent view of the challenges that test takers encounter when responding to a test item. Test taker verbal protocols (Buck, 1991; Vandergrift, 2007; Wu, 1998) might help us better understand reasons for potential item bias, by offering insights into how test takers of different age groups process listening comprehension items that demonstrate DIF. Until we can confidently uncover these processes, we have to stand with Nietzsche (2003, p. 139): “Against that positivism which stops before phenomena, saying ‘there are only facts,’ I should say: no, it is precisely facts that do not exist, only interpretations . . .”

ACKNOWLEDGMENTS

The authors contributed equally to this article and are listed alphabetically. The research reported in this article started when both authors were employed by CaMLA, the publisher of the Michigan English Test. Any opinions expressed in this publication are those of the authors and not necessarily of Educational Testing Service, the second author’s current affiliation. We thank colleagues at Educational Testing Service and CaMLA for their constructive feedback and advice on an early version of this article. We also thank the reviewers for their feedback, as a result of which the paper is much improved. Any flaws that remain are entirely our fault.

ORCID

Jayanti Banerjee http://orcid.org/0000-0002-8175-0887

REFERENCES

Abbott, M. L. (2007). A confirmatory approach to differential item functioning on an ESL reading assessment. *Language Testing*, *24*, 7–36. doi:10.1177/0265532207071510

Abdi, H. (2007). Binomial distribution/binomial and sign tests. In N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics* (pp. 88–90). Thousand Oaks, CA: Sage.

AERA, APA, NCME (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Alderson, J. C. (1990). Testing reading comprehension skills (Part 1). *Reading in a Foreign Language*, *6*, 425–438.

Alderson, J. C., & Kremmel, B. (2013). Re-examining the content validation of a grammar test: The (im)possibility of distinguishing vocabulary and structural knowledge. *Language Testing*, *30*, 535–556. doi:10.1177/0265532213489568

Allalouf, A., & Abramzon, A. (2008). Constructing better second language assessments based on differential item functioning analysis. *Language Assessment Quarterly*, *5*, 120–141. doi:10.1080/15434300801934710

Aryadoust, V. (2012). Differential item functioning in while-listening performance tests: The case of the International English Language Testing System (IELTS) listening module. *International Journal of Listening*, *26*, 40–60. doi:10.1080/10904018.2012.639649

Aryadoust, V., Goh, C. C. M., & Kim, L. O. (2011). An investigation of differential item functioning in the MELAB listening test. *Language Assessment Quarterly*, *8*, 361–385. doi:10.1080/15434303.2011.628632

Beglar, D. (2010). A Rasch-based validation of the vocabulary size test. *Language Testing*, *27*, 101–118. doi:10.1177/0265532209340194

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Brindley, G., & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing*, *19*, 369–394. doi:10.1191/0265532202lt236oa

Buck, G. (1991). The testing of listening comprehension: An introspective study. *Language Testing*, *8*, 67–91. doi:10.1177/026553229100800105

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

Cambridge Michigan Language Assessments (CaMLA). (2013). *MET 2013 report*. Ann Arbor, MI: CaMLA.

Cambridge Michigan Language Assessments (CaMLA). (2014). *MET 2009–2013 technical review*. Ann Arbor, MI: CaMLA. Retrieved from http://www.cambridgemichigan.org/about-us/research/

Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing*, *2*, 155–163. doi:10.1177/026553228500200204

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge, England: Cambridge University Press.

Elliott, M., & Wilson, J. (2013). Context validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening* (pp. 152–241). Cambridge, England: Cambridge English Language Assessment & Cambridge University Press.

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. doi:10.1080/15434300701375923

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. doi:10.1080/15434300701375758

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

Kim, M. (2001). Detecting DIF across the different language groups in a speaking test. *Language Testing*, *18*, 89–114. doi:10.1177/026553220101800104

Linacre, J. M. (2005). Rasch dichotomous model vs. One-parameter Logistic Model. *Rasch Measurement Transactions*, *19*, 1032.

Linacre, J. M. (2011). *WINSTEPS Rasch measurement computer program version 3.71.0*. Chicago, IL: Winsteps.com.

Linacre, J. M. (2014). *WINSTEPS Rasch measurement computer program version 3.81.0*. Chicago, IL: Winsteps.com.

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

McNamara, T. (1996). *Measuring second language performance*. Harlow, England: Longman.

Nietzsche, F. (2003). *Nietzsche: Writings from the late notebooks*. Cambridge, England: Cambridge University Press.

Pae, T.-I. (2004). DIF for examinees with different academic backgrounds. *Language Testing*, *21*, 53–73. doi:10.1191/0265532204lt274oa

Pae, T.-I., & Park, G.-P. (2006). Examining the relationship between differential item functioning and differential test functioning. *Language Testing*, *23*, 475–496. doi:10.1191/0265532206lt338oa

Papageorgiou, S. (2010). *Setting cut scores on the common European Framework of Reference for the Michigan English Test* (Technical Report). Ann Arbor, MI: University of Michigan.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.

Ryan, K. E., & Bachman, L. F. (1992). Differential item functioning on two tests of EFL proficiency. *Language Testing*, *9*, 12–29. doi:10.1177/026553229200900103

Sasaki, M. (1991). A comparison of two methods for detecting differential item functioning in an ESL placement test. *Language Testing*, *8*, 95–111. doi:10.1177/026553229100800201

Takala, S., & Kaftandjieva, F. (2000). Test fairness: A DIF analysis of an L2 vocabulary test. *Language Testing*, *17*, 323–340. doi:10.1177/026553220001700303

Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension research. *Language Teaching*, *40*, 191–210. doi:10.1017/S0261444807004338

Vandergrift, L., & Goh, C. C. M. (2012). *Teaching and learning second language listening: Metacognition in action*. New York, NY: Routledge.

Vandergrift, L., & Tafaghodtari, M. H. (2010). Teaching L2 learners how to listen does make a difference: An empirical study. *Language Learning*, *60*, 470–497. doi:10.1111/j.1467-9922.2009.00559.x

Wu, Y. (1998). What do tests of listening comprehension test? A retrospection study of EFL test-takers performing a multiple-choice task. *Language Testing*, *15*, 21–44. doi:10.1177/026553229801500102

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*, 223–233. doi:10.1080/15434300701375832

APPENDIX A

Example of a short conversation (occupational domain)

W1: Hi John, here's the summary of the budget report you asked me to write up.
M1: [surprised] Oh! Thanks Michelle, I wasn't expecting it to be completed so soon—fast work!
W1: Well, the report was pretty straightforward, so summarizing it took less time than I thought it would.

Why is the man surprised?

1. The summary has not been completed yet.
2. The summary was finished ahead of schedule.
3. The summary corrects a problem in the budget.
4. The summary raises questions about the budget.

Example of a short conversation (educational domain)

W2: You're taking biology with Professor Morrison, aren't you?
M2: Yeah . . . and it's really hard—lotsa homework . . . And his tests? Brutal. The bulk of my study time is devoted to his class.
W2: Hmm. I wanna take it next semester . . . but I'm gonna be taking some other difficult classes, too. Maybe I should hold off until next year . . .
M2: Definitely. If you're gonna have a heavy workload, you don't wanna add this to it.

What does the man recommend that the woman do?

1. talk to the professor
2. study with a partner
3. work only part-time
4. take the class next year

Example of a short conversation (public domain)

M1: Oh, no! I just remembered: I need to pick up my jacket from the cleaner today. I need it for my meeting tomorrow.
F1: You're not going to make it: They close early today.
M1: I guess I'll have to stop at the mall—I need a new one, anyway.

What is the man’s problem?

1. He won’t get to the cleaner on time.
2. He left his jacket at work.
3. He was late for work.
4. He doesn’t know where the mall is.