Language Assessment Quarterly

ISSN: 1543-4303 (Print) 1543-4311 (Online) Journal homepage: https://www.tandfonline.com/loi/hlaq20

An Investigation of Differential Item Functioning in the MELAB Listening Test

Vahid Aryadoust , Christine C. M. Goh & Lee Ong Kim

To cite this article: Vahid Aryadoust , Christine C. M. Goh & Lee Ong Kim (2011) An Investigation of Differential Item Functioning in the MELAB Listening Test, Language Assessment Quarterly, 8:4, 361-385, DOI: 10.1080/15434303.2011.628632

To link to this article: https://doi.org/10.1080/15434303.2011.628632

Language Assessment Quarterly, 8: 361–385, 2011

ISSN: 1543-4303 print / 1543-4311 online
DOI: 10.1080/15434303.2011.628632

Routledge
Taylor & Francis Group

An Investigation of Differential Item Functioning in the MELAB Listening Test

Vahid Aryadoust, Christine C. M. Goh, and Lee Ong Kim
National Institute of Education, Nanyang Technological University

Differential item functioning (DIF) analysis is a way of determining whether test items function differently across subgroups of test takers after controlling for ability level. DIF results are used to evaluate tests' validity arguments. This study uses Rasch measurement to examine the Michigan English Language Assessment Battery listening test for DIF across gender subgroups. After establishing the unidimensionality and local independence of the data, the authors used two methods to test for DIF: (a) a *t*-test uniform DIF analysis, which showed that two test items displayed substantive DIF, and favored different gender subgroups; and (b) nonuniform DIF analysis, which revealed several test items with significant DIF, many of which favored low-ability male test takers. A possible explanation for gender-ability DIF is that lower ability male test takers are more likely to attempt lucky guesses, particularly on multiple-choice items with unattractive distracters, and that having only two distracters makes this strategy likely to succeed.

This study evaluates the listening test of the Michigan English Language Assessment Battery (MELAB) through differential item functioning (DIF) analysis, a series of techniques that are used to uncover statistical bias in test items. The MELAB test was established in 1941 by the English Language Institute of the University of Michigan, as the Lado Test of Aural Comprehension, a multiple-choice listening test; this test evolved in 1951 into the Lado English Test, a multiple-choice grammar, vocabulary, and pronunciation test, and then into MELAB in 1985. According to the *MELAB Technical Manual* (Johnson, 2003), the MELAB test comprises listening, grammar, cloze, vocabulary, and reading sections, and an "impromptu composition."

Previous research has evaluated the MELAB listening test: Wagner (2004) and Eom (2008) studied the structure of the test, and the *MELAB Technical Manual* (Johnson, 2003) reports on its validity and reliability, including comparative data analyses of different subgroups of test takers. However, the test has not yet been subjected to DIF analysis (Weigle, 2000). Weigle argued that the test items of the MELAB test need to be examined to assess their fairness across various subgroups of test takers. Without such analysis, test users and researchers are left to presume that the test is fair and does not favor any group of test takers, an assumption that Wiegle's argument and previous studies (e.g., Goh & Aryadoust, 2010) suggest is likely inaccurate.

Test items display DIF when they function in favor of a subgroup of test takers. In other words, DIF occurs in test items when two groups of test takers with equal latent trait ability have

Correspondence should be sent to Vahid Aryadoust, English Language and Literature Academic Group, National Institute of Education, 1 Nanyang Walk, Singapore 637616. E-mail: ng0666907m@stdmail.nie.edu.sg

unequal probabilities of correctly answering a test item, a bundle of items, or even the entire test (Swaminathan & Rogers, 1990). Substantial observed DIF may represent evidence of bias and lack of fairness; interpreting observed DIF relies on expert judgment to provide a theoretical explanation (Ferne & Rupp, 2007). However, expert judgments have rarely been in line with statistical analyses (see Geranpayeh & Kunnan, 2007; Gierl & Khaliq, 2001).

This study focuses in particular on gender—because of its rather less ambiguous dichotomy, which is necessary for DIF analysis—as a potential cause of DIF; numerous studies have found that variation in certain test and item characteristics that are sensitive to gender differences (such as passage topic, item location, content, and vocabulary) may cause items to function differently across genders and cause gender-based DIF.[1] For example, in two early studies into the effect of gender on test performance, Ferber, Birnbaum, and Green (1983) and Lumsden and Scott (1987) reported that female test takers outperformed male test takers in essay items, whereas male participants were better at multiple-choice questions (MCQs). They proposed that male individuals might be more knowledgeable, but they did not explain why this might be. Likewise, Walstad and Robson (1997) showed that MCQs favor male test takers in economics tests and hypothesized that the observed DIF might reflect gender differences in cognitive processes, as well as "differential reasoning, socialization, instructional practices, or the format used for testing" (p. 168).

Breland, Lee, Najarian, and Muraki (2004) found that open-ended questions—questions where the test taker must supply, not choose, the correct answer—generally favored female test takers, which they hypothesized might have been due to gender differences in reasoning and cognitive processes. Lin and Wu (2003), analyzing test data from an English test modeled after the paper-based version of the Test of English as Foreign Language (TOEFL®) and administered to Chinese test takers, found that the test's 30-item MCQ listening comprehension section clearly favored female test takers, contradicting Ferber et al.'s (1983) and Lumsden and Scott's (1987) predictions, as well as studies by Boyle (1987) and Ryan and Bachman (1992), which found no substantive DIF in the TOEFL®.

More recently, Aryadoust (in press), investigating DIF in the International English Language Testing System (IELTS™) listening test, found that low-ability male test takers performed unexpectedly well on MCQs with only three options from which to choose. Drawing on the erratic response patterns of low-ability male individuals, Aryadoust hypothesized that male participant were more likely to attempt lucky guesses on difficult items and that the high probability of a lucky guess with only three choices rewarded this behavior. Although much has been learned from these studies, it is as yet unknown what provokes the observed variation in gender-based DIF between tests. MCQs seem to favor different genders under different circumstances, likely reflecting gender-based differences in psychological processes.

Language assessment researchers have begun using statistical DIF analysis (principally Rasch-based methods) comparatively recently (see, e.g., Muraki, 1999; Roznowski & Reith, 1999; Ryan & Bachman, 1992; Takala & Kaftandjieva, 2000; Zenisky, Hambleton, & Robin, 2003; Zhang, Matthews-Lopez, & Dorans, 2003). Rasch-based studies typically use Rasch mean square (MNSQ) and standardized fit statistics to assess the applicability of the data set to the model (for additional details, see The Rasch Model section). Discrepant fit indices can indicate

[1] We are thankful to an anonymous reviewer for making this comment.

erratic patterns in data, likely due to cheating, carelessness, or lucky guesses (Linacre, 2010a). However, no effective fit range has been established. Some researchers (Bond & Fox, 2007) have proposed a liberal range, with MNSQs between 0.6 and 1.4. Others (Wright & Linacre, 1994) have proposed a far more stringent range, with MNSQs between 0.8 and 1.2. Language assessment researchers have typically applied the more liberal range and have identified test items as "fitting the model," when the application of the more stringent range would have identified the items as exhibiting DIF (Jang & Roussos, 2007, 2009; Mazor, Clauser, & Hambleton, 1994).

Many studies have also failed to test for nonuniform DIF (NUDIF), which occurs when a test item favors different subgroups at different levels of trait ability (e.g., favors low-ability male participants over low-ability female participants, and high-ability female participants over high-ability male participants). Several studies that have not detected UDIF have been shown to have NUDIF bias in their test items (see Mazor et al., 1994).

This study uses Rasch-based DIF analysis to examine gender-based DIF in data from 852 takers of Form FF of the MELAB listening test. The study's objectives were to determine

1. whether the test data support the assumptions of unidimensionality and local independence, as preconditions for Rasch-based DIF analysis;
2. whether there exists evidence of substantive gender-based DIF in the MELAB listening test, and whether the observed DIF is a function of ability (i.e., is nonuniform);
3. the factors that cause DIF in items in the test; and
4. whether more stringent Rasch fit criteria can indicate the presence of DIF.

DIF ANALYSIS

For a test item to display DIF implies a persistent interaction between the performance of a subgroup of test takers and an attribute (e.g., age, gender, race, or nationality), which would give an unfair advantage to that subgroup over another (see Kunnan, 1990; Zeidner, 1986, 1987). To meaningfully impact test scores, this interaction must be not only too improbable to be attributable to chance but substantive as well. Statistically significant DIF indices may nevertheless be too small in magnitude to have any meaningful effect on the measurement (Linacre, 2010a). Therefore, to cause test bias, DIF must be statistically significant ($p < .05$), substantively impact observed test or test item performance, and have a theoretically sound cause.

Significant and substantive DIF indices imply that test scores no longer represent only the intended latent variable; they also represent an unintended and unmodeled secondary dimension (Wright & Stone, 1988). Unmodeled secondary dimensions may be either simple or complex (Jang & Roussos, 2009). The presence of a simple secondary dimension indicates that most test items measure the intended trait but that a group of items measures a secondary attribute that is nevertheless targeted on the intended trait. These items form an "auxiliary dimension" (Jang & Roussos, 2009, p. 242). The presence of a complex secondary dimension means that test items measure unintended traits, whose degree and type differs from item to item (Jang & Roussos, 2007, 2009). In tests with complex secondary dimensions, test items have primary and auxiliary dimensions, which measure the latent trait, and at least one "nuisance dimension," which does not (Jang & Roussos, 2009, p. 242). Ackerman, Gierl, and Walker (2003) referred to DIF caused by auxiliary dimensions as *benign* and that caused by nuisance dimensions as *adverse*.

This study adopts Rasch-based DIF analysis, one of the most frequently used methods of DIF analysis. Wyse and Mapuranga (2009) argued that the Rasch method is broadly comparable to other methods, and Cauffman (2006) and Edelen, McCaffrey, Marshal, and Jaycox (2009) have reported on the potential of the method to detect gender-based DIF in educational assessment. The Rasch model enjoys the important advantage of being able to detect both uniform DIF (UDIF) and NUDIF (Linacre, 2010a). Most other models with the exception of logistic regression (Swaminathan, 1994) can detect only the former. Rasch-based DIF analysis has two preconditions: (a) unidimensionality, which holds when overall test scores are not contaminated by any irrelevant factor, and (b) local independence, which holds when test takers' performance on a given test item is not influenced by their performance on another item (Ferne & Rupp, 2007).

Dimensionality analysis and DIF analysis are conceptually distinct. Dimensionality analysis yields information about secondary dimensions that are relevant to *all* test takers, whereas DIF analysis identifies conditional differences in response probabilities using defined variables (such as gender) that dimensionality analysis does not examine.[2] In a large test, a few items may exhibit marked DIF, which will nevertheless be undetectable in dimensionality analysis if it is not substantial enough to explain a large amount of the overall variance in responses. Roussos and Stout (1996, 2004) argued that although the presence of DIF points to multidimensionality, "the presence of a secondary dimension does *not* automatically imply the presence of DIF. Some secondary dimensions cause DIF and some do not, depending on how the reference and focal groups differ in their proficiency on the secondary dimension" (Roussos & Stout, 2004, p. 108). Because of these distinctions, dimensionality analysis is an important precondition to Rasch-based DIF analysis (Ferne & Rupp, 2007, p. 129). Unfortunately, only eight of 27 studies in Ferne and Rupp's survey of DIF analysis in language assessment provided evidence of unidimensionality. However, the MELAB listening construct has previously been shown to be statistically unidimensional (Goh & Aryadoust, 2010; Liao, 2007; A. Wagner, 2004).

DIF analysis requires that researchers distinguish two subgroups: reference and focal. The reference subgroup's performance is believed likely to match researchers' expectations, and the focal subgroup's performance is believed likely to exhibit DIF (Luppescu, 1993). This distinction is made because DIF analysis is often employed to identify DIF suggested by theory. For example, if learners of a language join native speakers in taking a chemistry test written in the language they are learning, the native speakers form the reference group and the language learners form the focal group.

As previously discussed, DIF can be classified as either UDIF or NUDIF (Ferne & Rupp, 2007). UDIF indicates that the subgroup differences in the secondary dimension are constant across the main dimension and that "there is no interaction between ability level and group membership" (Prieto Maranon, Barbero Garcia, & San Luis Costas, 1997, p. 559). This implies that the item characteristic curves (ICCs) of two subgroups have identical slopes but different intercepts, indicating a consistent difference across the two subgroups (e.g., male and female), irrespective of the subclass being examined (e.g., low- or high-ability test takers). NUDIF, conversely, does vary with the ability level of test takers. In other words, the difference in performance between two subgroups is not consistent between subclasses of those subgroups. In NUDIF, group membership interacts with ability levels to form "nonparallel item

[2]We are thankful to Mike Linacre and André Rupp for their comments on dimensionality analysis.

characteristics curves" (Prieto Maranon et al., 1997, p. 559); because their slopes differ, these ICCs intersect (Zumbo, 1999)—usually at the mean in a standardized sample—and the test item switches from favoring one subgroup to the other. If NUDIF acts against a subgroup it is called "negative DIF," and if it favors a subgroup it is called "positive DIF" (Camilli & Shepard, 1994, pp. 59–60). Positive and negative DIF magnitudes may balance. Investigating NUDIF is an important concern, because it is often muted in research and many studies that have not detected UDIF have been shown to have NUDIF bias in their test items (see Mazor et al., 1994). Recently, Ferne and Rupp (2007) argued that "failure to consider the possibility of nonuniform DIF can have serious practical implications" (Ferne & Rupp, 2007, p. 134). For example, developing efficient cutoff scores to make significant academic and educational decisions "about otherwise disadvantaged student groups" will be compromised in the absence of NUDIF analysis results (Ferne & Rupp, 2007, p. 134). It is therefore important that the chosen model for analysis have the ability to detect both UDIF and NUDIF which the Rasch model, like logistic regression (Swaminathan, 1994), can do unlike other models (Linacre, 2010a).

We can place the present study in the framework of the "second DIF generation" proposed by Zumbo (2007). Zumbo summarized the approaches to DIF analysis into three trends. The first generation is marked by the use of the term "item bias" analysis, dichotomous items, and focal/reference groups. The second generation is signaled by the common use of the term *DIF* in lieu of item bias, item response theory, and regression models. Within this trend, researchers focus on UDIF, NUDIF, and dimensionality analysis. The third generation goes beyond the multidimensionality in DIF studies and is marked by the multiple-indicators, multiple causes structural equation modeling. This framework recognizes the potential influence of "contextual variables such as classroom size, socioeconomic status, teaching practices, and parental style" on DIF (Zumbo, 2007, p. 229).

Although DIF analysis has expanded considerably in sophistication over the past decade, and although a number of studies have been conducted on DIF in language assessment, the literature still lacks a solid theory of the means of investigation of DIF (see Zumbo, 2007). Researchers may take either an exploratory approach, in which they perform a post hoc content analysis of those items that display DIF (e.g., Lin & Wu, 2003), or a confirmatory approach, in which they analyze test items to generate hypotheses, which they then test through DIF analysis (e.g., Gierl, 2005). In an extensive review of DIF analysis in language testing comprising 27 studies, Ferne and Rupp (2007) found that most researchers used exploratory analysis. Ferne and Rupp pointed out that confirmatory analysis is generally preferred because it can provide supporting or attenuating evidence for a postulated theory of DIF. However, one advantage of exploratory analysis is that it can lead to new theories of DIF. In either type of analysis, researchers need a solid framework to judge their findings and propose causes of observed DIF, but attempts to identify these causes are often unsuccessful, and many studies leave the presence of DIF as a mystery to stimulate further research (Ferne & Rupp, 2007).

LINGUISTIC CAUSES OF DIF

Because the literature on DIF in listening comprehension tests is critically narrow, we draw on findings from general DIF studies and studies investigating other language skills where applicable. Uiterwijk and Vallen (2005) classified the linguistic causes of DIF in reading tests into four

TABLE 1
Levels and Causes of Differential Item Functioning (DIF) in Reading Tests

Levels of DIF Cause	*Remarks and Details*
Word	Meaning; frequency; abstractness; ambiguity.
Sentence	Negative/positive; metaphors; idioms.
Text	Requiring memorization; implausible content; complexity in referencing; strange or wrong clues in the text structure.
Metalinguistic competence	Items ask to identify grammatically ill-formed structures; items ask to make a correction.

Note. This table is a summary of Uiterwijk and Vallen's (2005) theoretical framework.

major categories: word, sentence, text, and metalinguistic competence level (see Table 1). This framework is informed by previous research by Coenen and Vallen (1991), Uiterwijk (1994), and Uiterwijk and Vallen (1997). Uiterwijk and Vallen (2005) used their framework to compare the performance of second-generation immigrants and native speakers on a Dutch language battery test. They hypothesized that the framework would explain observed DIF in their study, but this did not happen. Similarly, Mahoney (2008) investigated the effect of language complexity (operationalized as word length, modifier locations, use of passive voice, and morphology) on DIF in a math test and found that the factors she tested did not cause DIF for either first- or second-language test takers.

More recently, Aryadoust (in press) investigated DIF in the IELTS™ listening test and found that items that use negative morphemes favor high-ability test takers. This finding agrees with recent findings in linguistics indicating that negative statements appear to be formed and understood in a number of cross-culturally different ways (Horn, 2010) and that language learners often struggle with the "formal expression" of negative forms in the new language (Dimroth, 2010, p. 40).

These theoretical attempts to identify the causes of DIF have been informative but relatively few in number. Numerous DIF studies have focused on conducting statistical analysis rather than testing theoretical predictions. Neither the exploratory nor the confirmatory approach has been successful in developing and refining a framework for explaining DIF, because the theoretical interpretations of DIF tend to be incongruent with the statistical findings. Roussos and Stout (1996) stated that "attempts at understanding the underlying causes of DIF using substantive analyses of statistically identified DIF items have, with few exceptions, met with overwhelming failure" (p. 360). This concern was shared by the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) concluding that "there has been little progress in identifying the causes or substantive themes that characterize items exhibiting DIF" (p. 78). More recently, Roussos and Stout raised the same concern again. Reviewing the DIF literature, they found that

> unfortunately, throughout most of the history of DIF statistics, they have been used in almost total isolation from the substantive aspects. Until recently, the standard DIF procedure has been the application of a DIF statistic in an automatic one-item-at-a-time purely statistical analysis. (Roussos & Stout, 2004, p. 108)

The lack of a solid theoretical DIF framework to guide researchers toward identifying the substantive causes of statistical DIF may account for such "automatic" statistical analyses.

LISTENING COMPREHENSION AND THE MELAB LISTENING CONSTRUCT

Although no consensus exists on the underpinning structure of the listening construct, research shows that a variety of abilities contribute to listening ability, such as vocabulary knowledge (Maccartty, 2000), visuals (A. Wagner, 2004), accent (Derwing & Rossiter, 2003), grammar knowledge (Liao, 2007), and recall ability (Klaassen & Snippe, 1998).

To define a listening construct, it is crucial to distinguish between listening comprehension "prerequisites" such as attention and recognition, comprehension itself, and its applications (Dunkel, Henning, & Chaudron, 1993). Although there may be a significant correlation between prerequisites, applications, and comprehension, items that merely get at the former two might not tap the construct. Flowerdew and Miller (2010) proposed that the facets of an "eclectic" listening construct include bottom-up, top-down, and interactive processing (p. 167). Bottom-up processing entails piecing together smaller units, such as morphemes, to reconstruct larger elements, such as words, phrases, and sentences. In top-down processing, listeners rely on their knowledge or schemata to make sense of the message. Interactive processing draws on both bottom-up and top-down processes simultaneously (see Goh, 2005, 2010, for a review).

The *MELAB Technical Manual* (Johnson, 2003) summarizes the MELAB listening construct as the ability to (a) use schemata to interpret meaning; (b) use components of one's linguistic system, such as grammar and vocabulary, to construct understanding; (c) use a range of comprehension skills and strategies; and (d) make inferences and draw conclusions. This construct representation generally resonates with Flowerdew and Miller's (2010) definition, as it embraces both bottom-up and top-down processes and their interactions. Yet a close examination of the test seems to show that approximately one third of items are highly decontextualized. Given the presence of these items, Weigle (2000) argued that the interactivity of the test is not well established.

Although the MELAB test's reliance on decontextualized, discrete-point items increases its reliability (Weigle, 2000), this design purportedly compromises some aspects of test usefulness (Buck, 2001; Weigle, 2000): "Listening to a series of unrelated, decontextualized questions and statements, even if the micro-skills involved in these tasks may be relevant for academic writing and listening," limits the test's authenticity and interactivity (Weigle, 2000, p. 451).

METHODOLOGY

Data Source

The test data used in the study were provided by the English Language Institute of the University of Michigan. The data set comprised the performance of 852 test takers on 50 listening test items. Test takers were from 78 countries across the world; 425 (49.9%) were female and 427 (50.1%) were male. Following previous studies in language assessment, we selected the male subgroup as the reference group and the female subgroup as the focal group.

Materials

The MELAB listening test comprises 50 MCQ items delivered to test takers aurally and divided into three sections: Section 1, 15 minimal-context items; Section 2, 20 short conversation items; and Section 3, three long conversation prompts, each with five corresponding test items. Tests are scored on a scale from 30 to 100, with an average score of 80 (Johnson, 2003). The test data in this study were from an administration of Form FF of the test, which is a secure test version.

Data Analysis

Prior to undertaking DIF analysis, we performed two major analyses of the test data: (a) an examination of its descriptive statistics, item difficulty measures, fit to the Rasch model, and reliability, and (b) a test of its dimensionality and degree of local independence.

Descriptive statistics. We calculated descriptive statistics for the test data, including mean, standard deviation, and skewness and kurtosis coefficients, using SPSS for Windows, Version 16 (SPSS Inc., Chicago, IL).

The Rasch model. In the logistic Rasch model for dichotomous data, a test item's difficulty measure is calculated from the proportion of all test takers who answer it correctly, irrespective of those test takers' ability levels; a test taker's ability measure is calculated from the total number of items he or she answers correctly, irrespective of their difficulty. The Rasch model creates a logistic function based on the difference between person ability and item difficulty. The wider the difference between ability and difficulty "in favor of" the person, the more likely that person is to answer the item correctly. That is, the greater the ability of the person, relative to the difficulty of the item, the more likely that person is to answer the item correctly (Wright & Stone, 1988).

We calibrated test items on the Rasch model using WINSTEPS, Version 3.69 (Linacre, 2010b). As part of Rasch analysis, we calculated infit and outfit mean square (MNSQ) indices for all test items. Infit MNSQ is an information-weighted inlier-sensitive index sensitive to departures from expected patterns in test items closer to average difficulty (Linacre & Wright, 1994, p. 360), and outfit MNSQ is an outlier-sensitive index (Linacre, 2002) sensitive to departures from expected patterns in test items of low or high difficulty. Of the two, the infit index is more commonly reported in measuring the fit of data to the Rasch model because it captures erratic patterns in the data near the ICC. MNSQ is calculated by dividing test items' chi-square statistics by their degrees of freedom (Wright & Linacre, 1994). The expected MNSQ value is 1.0, so a value of 1.1 has 10% more noise than expected by the model, which increases the standard error of measurement (Wright & Linacre, 1994). Bond and Fox (2007) defined an item as underfitting when its MNSQ indices are greater than 1.4 and as overfitting when they are less than 0.6. However, Wright and Linacre (1994) proposed a more stringent fit criterion that regards items falling out of the region between 0.8 and 1.2 as misfit. Wright and Linacre's stringent criterion is preferred in this context because it seems to suit dichotomous data better (Smith, 1996) than Bond and Fox's.

We also took point-measure correlations for all test items. These correlations measure the agreement of observed scores with the latent trait. As is commonly done in Rasch analysis, we

display the relationships between persons and items on an item-person map (or Wright map, in honor of Benjamin Wright), which arrays both person performance and item difficulty along a single line calibrated in log-odd units (logits).

Reliability analysis. We tested the reliability of the data using Rasch model analysis. In Rasch analysis, reliability is estimated for both persons and items and ranges from 0 to 1. A test's person and item reliability indices are a measure of the precision with which the test measures person performance and item difficulty (Linacre, 2010a). Low reliability indicates that the variability in measures is contaminated by a high standard error of measurement (SEM). SEM may inflate if test takers do not get enough items corresponding to their ability, so shorter tests have lower reliability coefficients.

Reliability is also expressed as another index known as separation, which is the ratio of test items' or test takers' standard deviation to their root mean square standard error (Linacre, 2010a), and ranges from zero to infinity.

Unidimensionality and local independence. We used Rasch measurement to test for both unidimensionality and local independence, as prerequisites to DIF analysis.

We tested for unidimensionality using principal component analysis of linearized Rasch residuals (PCAR). Because items do not usually match the Rasch model perfectly, some residual remains after fitting the data to the Rasch model. This residual is the difference between the expectations of the Rasch model and the observed data (Linacre, 1998a; Wright, 1996a). PCAR can identify components (or secondary dimensions) in residuals by estimating their common variance. If some item residuals have substantially high common variance, they cluster together and construct a component. If an emergent component is observed, its magnitude (i.e., its ability to explain the common variance in data) should be compared with that of the Rasch model. If the component accounts for three or more items (or units) out of all test items (i.e., three items are distinguishable as a cluster), and if the difference is large in favor of this component, then the component is likely substantial and the assumption of unidimensionality is attenuated (Linacre, 2010a); otherwise, the common variance cannot be attributed to a variable other than those the test taps into, and unidimensionality is supported (Linacre, 1998a).

To carry out PCAR, we initially undertook a Rasch analysis of the test data to form a linear measure, followed by a conventional PCA of the linearized residuals. The results are illustrated in the appendix as a graph plotting the item difficulty measures against the magnitude of component loadings of their linearized residuals. The items that have higher loading magnitudes land at the top or bottom of the graph, indicating that a substantial part of their variance is left unmodeled or unexplained by the Rasch dimension. If items with residual loading magnitudes significantly above zero cluster together on the graph, this merits a closer investigation into the substance and theoretical implications of the cluster. No cluster of items is distinguishable from the graph in the appendix, and the first extracted component explains the common variance in only 1.7 items (2.9%), which is not substantial.

PCAR enjoys a number of advantages over other dimensionality analysis methods such as exploratory factor analysis (EFA): (a) EFA factors are not necessarily dimensions; they can be clusters of items correlating substantially due to their difficulty measures (Linacre, 1998b; Wright, 1994a, 1994b); (b) results of EFA are often obscured by "ordinality" of variables and high correlations among factors (Schumacker & Linacre, 1996, p. 470; Smith, 1996), whereas

linearized data do not generate "illusory factors" (Linacre, 1998b, p. 603); (c) missing data can bias the EFA factor solution, but Rasch measurement is resilient to missing data (Wright, 1996b); and (d) factor analysis stratifies the construct, whereas Rasch analysis spans the data "with one meaning" (Bond, 1994, p. 374).

We also calculated fit statistics, which can hint at multidimensionality. Test items displaying erratic fit indices have inaccurate estimated difficulty measures and are likely to be contaminated by a factor not intended by the test designer.

We tested for local independence using Pearson correlation analysis of linearized Rasch residuals. Residuals, the difference between the observed difficulty measure of items and the values expected by the Rasch model, indicate "how much locally easier or harder that item was than expected" (Wright, 1994b, p. 510). Local independence holds when no datum influences another and is evidenced by small absolute values of correlations between item residuals.

DIF. We calculated the DIF effect size of each test item. This statistic is the ratio of the contrast in local item difficulty between subgroups to the standard deviation of the reference group. Effect size is negligible if it is below 0.4, "slight to moderate" if it approximates 0.4, and "moderate to large" if it is above 0.6 (Linacre, 2010a, p. 487). We also contrasted local difficulty measures by gender using a Welch t test. We followed Linacre (2010a) in using a p value of .05 ($p < .05$) for this test.

We used the Rasch model to investigate gender-based UDIF and NUDIF. Once UDIF has been found, the researcher should investigate whether it is a statistical or substantive case by further dividing subgroups into high- and low-ability subsections and conducting NUDIF analysis. DIF that persists in this analysis represents evidence of systematic DIF (Du, 1995). As Du (1995) stated,

> The best [confirmatory] test for DIF is "Does it replicate?" When the focal group is split in various ways, random or systematic, does the item continue to exhibit DIF in the same way for each subgroup? If so, the DIF shows evidence of being real. If not, it may be just an accident of sampling. (p. 414)

When DIF is statistically and substantively significant and replicates across various subgroups, the researcher can more confidently regard the item as functioning differently across subgroups.

RESULTS

The prime objectives of this article include the evaluation of unidimensionality and local independence in the MELAB listening test, as preconditions for Rasch-based DIF analysis, gender-based UDIF and NUDIF, factors that cause DIF in items, and whether more stringent Rasch fit criteria can indicate the presence of DIF. Our findings are discussed, as follows.

Fit of the Data to the Latent Trait Model

Table 2 presents descriptive statistics for the test data, as well as Rasch measurement results, including difficulty measures in logits, fit indices, and point-measure correlations.

TABLE 2
Results of Descriptive Statistics Analysis and Rasch Measurement

Descriptive Statistics					*Rasch Measurement*				
Item	*M*	*SD*	*Skewness*	*Kurtosis*	*Measure*	*Infit MNSQ*	*Outfit MNSQ*	*PT-Measures*	*Total Scores*
1	0.48	0.500	0.08	−1.98	0.76	1.05	1.04	0.33	408
2	0.52	0.500	−0.08	−1.98	0.57	1.04	1.05	0.32	444
3	0.63	0.482	−0.56	−1.68	0.03	1.01	1.02	0.32	541
4	0.74	0.438	−1.10	−0.77	−0.54	0.93	0.85	0.37	632
5	0.70	0.460	−0.85	−1.27	−0.28	0.89	0.83	0.42	593
6	0.45	0.498	0.18	−1.97	0.88	1.13	1.13	0.25	387
7	0.59	0.493	−0.34	−1.88	0.26	0.96	0.94	0.39	499
8	0.87	0.338	−2.18	2.78	−1.44	0.95	0.84	0.28	740
9	0.80	0.397	−1.53	0.35	−0.93	0.91	0.82	0.35	685
10	0.80	0.403	−1.47	0.18	−0.88	0.95	0.86	0.33	679
11	0.74	0.439	−1.09	−0.80	−0.52	1.01	1.00	0.28	630
12	0.48	0.500	0.06	−2.00	0.74	1.03	1.03	0.34	413
13	0.55	0.497	−0.21	−1.95	0.42	1.05	1.03	0.31	471
14	0.54	0.499	−0.14	−1.98	0.50	1.03	1.01	0.33	456
15	0.75	0.433	−1.16	−0.64	−0.59	0.95	0.92	0.34	640
16	0.54	0.499	−0.15	−1.98	0.49	1.03	1.02	0.33	458
17	0.50	0.500	−0.01	−1.89	0.65	1.06	1.05	0.31	429
18	0.60	0.491	−0.39	−1.84	0.21	0.98	0.96	0.37	509
19	0.60	0.490	−0.41	−1.83	0.19	0.96	0.91	0.39	512
20	0.61	0.488	−0.45	−1.79	0.15	1.03	1.05	0.31	520
21	0.54	0.498	−0.17	−1.97	0.46	1.01	0.99	0.35	463
22	0.62	0.486	−0.48	−1.76	0.11	0.98	0.92	0.37	527
23	0.67	0.472	−0.70	−1.50	−0.13	1.01	0.97	0.32	568
24	0.79	0.407	−1.43	0.05	−0.84	0.89	0.76	0.39	674
25	0.68	0.466	−0.78	−1.39	−0.21	1.02	1.02	0.29	581
26	0.40	0.490	0.41	−1.83	1.14	1.16	1.22	0.22	340
27	0.51	0.500	−0.03	−2.00	0.62	1.03	1.07	0.33	434
28	0.67	0.470	−0.73	−1.46	−0.16	1.02	1.03	0.29	573
29	0.59	0.491	−0.38	−1.85	0.22	0.95	0.91	0.40	506
30	0.69	0.462	−0.83	−1.31	−0.26	0.89	0.79	0.44	589
31	0.63	0.484	−0.52	−1.73	0.07	0.96	0.94	0.37	533
32	0.69	0.463	−0.82	−1.32	−0.25	0.95	0.94	0.36	588
33	0.75	0.431	−1.18	−0.59	−0.61	0.96	1.11	0.31	643
34	0.76	0.430	−1.19	−0.57	−0.62	1.02	1.04	0.27	644
35	0.63	0.483	−0.54	−1.71	0.05	0.98	0.95	0.36	537
36	0.52	0.500	−0.07	−2.00	0.58	1.19	1.25	0.17	441
37	0.67	0.469	−0.74	−1.45	−0.17	0.93	0.85	0.40	574
38	0.54	0.498	−0.17	−1.97	0.46	1.02	1.01	0.34	463
39	0.61	0.487	−0.46	−1.78	0.13	1.02	1.01	0.32	522
40	0.57	0.495	−0.29	−1.91	0.32	1.03	1.01	0.33	488
41	0.62	0.487	−0.47	−1.77	0.12	1.04	1.10	0.28	525
42	0.69	0.462	−0.83	−1.30	−0.27	0.97	0.97	0.34	590
43	0.79	0.411	−1.39	−0.06	−0.80	0.92	0.78	0.37	669
44	0.50	0.500	0.01	−2.00	0.69	1.05	1.11	0.31	422

(*Continued*)

TABLE 2
(*Continued*)

Descriptive Statistics					*Rasch Measurement*				
Item	*M*	*SD*	*Skewness*	*Kurtosis*	*Measure*	*Infit MNSQ*	*Outfit MNSQ*	*PT-Measures*	*Total Scores*
45	0.84	0.370	−1.82	1.33	−1.17	1.02	1.09	0.21	713
46	0.77	0.423	−1.26	−0.38	−0.69	0.91	0.83	0.37	654
47	0.64	0.480	−0.58	−1.66	0.00	1.02	1.05	0.31	545
48	0.67	0.469	−0.74	−1.45	−0.17	1.06	1.15	0.25	574
49	0.47	0.500	0.10	−1.99	0.79	1.02	1.08	0.31	403
50	0.65	0.477	−0.63	−1.59	−0.06	0.98	0.95	0.35	555

Note. $n = 852$. MNSQ = mean square; PT-measures = point-measure correlations.

Item 8 ($M = 0.87$, total score = 740) has the highest mean score, and Item 26 ($M = 0.40$, total score = 340) the lowest, indicating that Item 8 was answered correctly and Item 26 incorrectly by many test takers. Item 8 was the easiest and Item 26 the most difficult. Skewness and kurtosis coefficients fall between –2 and +2 in all items except Item 8, indicating univariate normality.

The Infit MNSQ and Outfit MNSQ columns in Table 2 present the test items' infit and outfit MNSQ indices. All fit statistics fall within the range from 0.6 to 1.4 recommended by Bond and Fox (2007), but some (Items 24, 26, 30, 36, and 43—10% of all items) fall out of the range from 0.8 to 1.2 recommended by Wright and Linacre (1994). Because there is no established Rasch fit standard for language tests, we could not decide whether Bond and Fox's more liberal criterion functions better than Wright and Linacre's more stringent criterion at this stage, although we considered this later in the DIF analysis. If we regard the stringent criterion as having a better explanatory power, it exceeds the 5% rate expected to occur by chance, which is noteworthy in the sense that the observed DIF cannot be attributed to chance.

Several items, such as Items 1, 2, and 3, show MNSQ fit indices near 1, indicating a lack of erratic response patterns in the data. Some test items, such as Items 4 and 5, display MNSQ fit statistics below 1 and overfit the model to some degree, and some, such as Items 26 and 36, underfit to some degree, indicating unexpected variance, possibly due to guessing or carelessness (Wright & Linacre, 1994). The PT-Measures column gives point-measure correlations for test items; all correlations are positive. Together, these two findings indicate that many observed scores accord with the expectations of the Rasch model.

Figure 1 plots person and item locations concurrently (i.e., how items spread out in relation to the ability of test takers). The leftmost column gives the measurement scale in log-odd units (logits). Each hash mark (#) on the left side of the dividing line represents six test takers, and each number preceded by "v" (variable) on the right represents the test item of that number. On each side of the dividing line, *M* represents the mean, *S* is 1 standard deviation from the mean, and *T* is 2 standard deviations from the mean. Test items range from easiest at the bottom to most difficult at the top, and test takers range from lowest scoring at the bottom to highest scoring at the top.

The map shows that test items cover a relatively wide range of difficulty, from –1.44 logits (Iitem 8; *SEM* = 0.10) to +1.14 logits (Item 26, SEM = 0.08), with an even spread. Excluding 56 people with measures greater than 2.00 logits (from +2.11 to +5.26), person ability measures

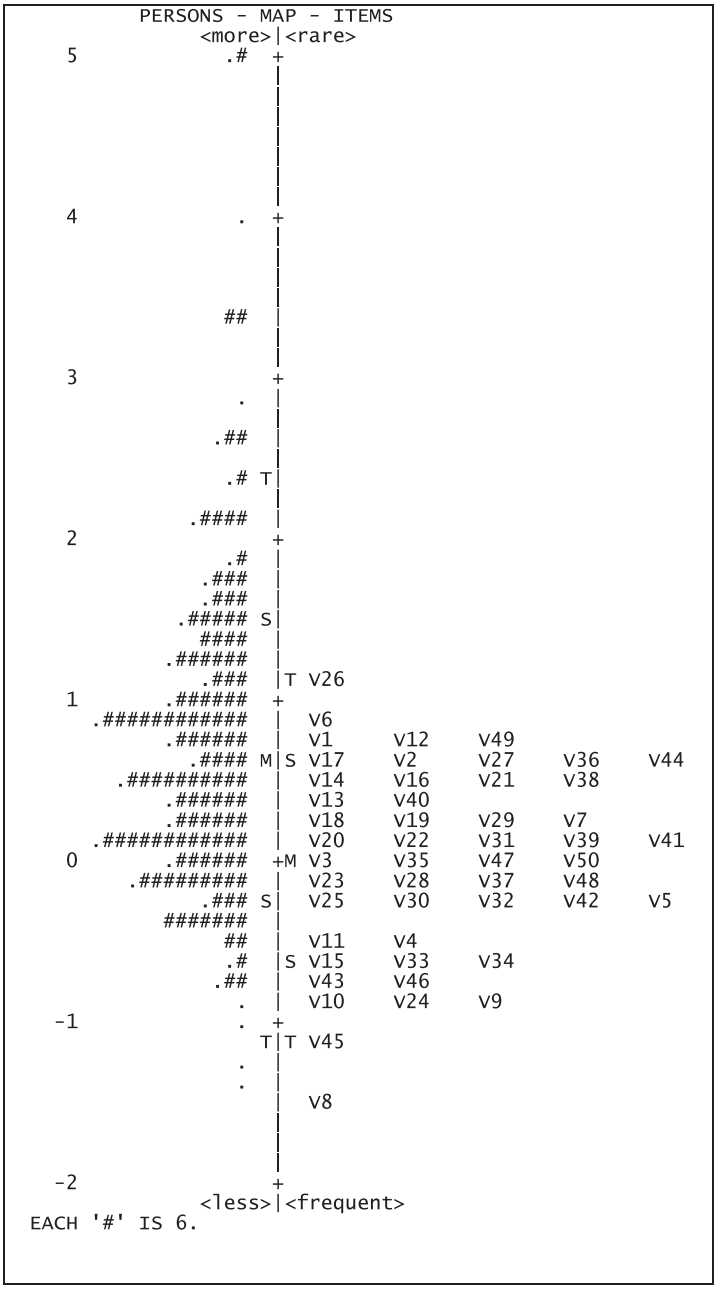


FIGURE 1 Calibration of items in the test items.

range from –1.35 (SEM = 0.44) to +1.93 (SEM = 0.35). It is not unexpected to observe some high-ability test takers with no test items in their ability levels. They have already demonstrated their ability to pass the test, whose important objective is to distinguish able test takers from those with minimum listening proficiency. Therefore, test items center around the mean, where the majority of test takers land. In addition, no gaps are observed in the item hierarchy, and most items are similar in difficulty to a number of other items, indicating that there are sufficient items to estimate test takers' ability, especially around the mean where the majority of test takers landed.

Rasch Reliability Analysis

The item reliability coefficient is 0.98 and the person reliability coefficient is 0.88, meaning that only 2% of the variability in item measures and 12% of variability in person measures is attributable to error. Item and person separations are 2.33 and 7.17, respectively, indicating that this measurement consistently discriminates approximately two statistically distinct strata of performance in persons and seven difficulty levels in items (Wright, 1996b).

Unidimensionality and Local Independence

PCAR shows that the Rasch model explains 31% of the observed variance, and the first component in the residuals only 2.5%. As the appendix displays, the item residuals do not form distinguishable clusters, indicating that they contain no substantive (i.e., theoretically meaningful) structure. This observation is backed by statistical outputs: Because the extracted dimension from the residuals is 12.5 times smaller than the Rasch dimension, it is not substantive (Linacre, 2010a), supporting the assumption of unidimensionality.

Next, we analyzed the Pearson correlations, which strongly supported the assumption of local independence. Correlations above .70 are evidence of local dependence (Linacre, 2010a), and all observed correlations fell between zero and .15.

Identification of Differential Item Functioning

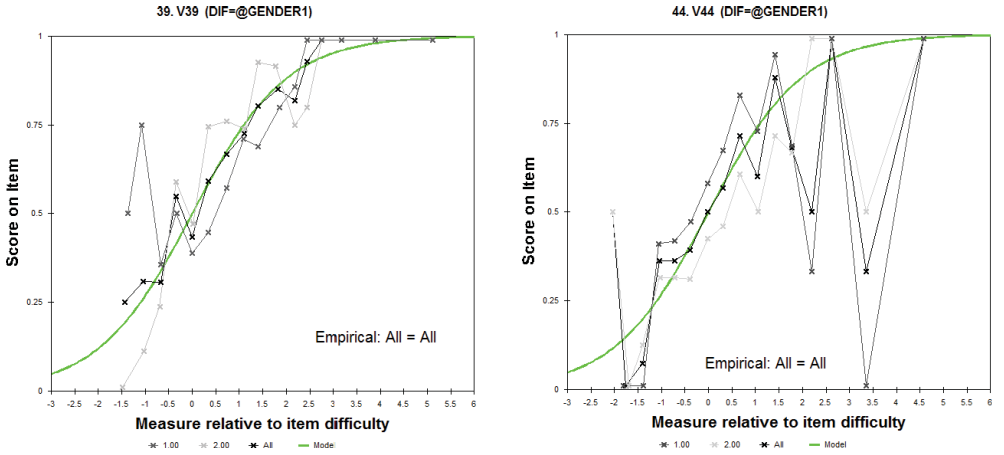
Table 3 displays a DIF analysis of all test items, including the local difficulty of test items for each gender subgroup, SEM figures for each measurement, the local difficulty contrast between gender subgroups, and a Welch *t* value and a *p* value for this contrast. The Welch *t* value presents the statistical difference between the local difficulties of items as a Student's two-sided *t* statistic (Linacre, 2010a). Reading across Table 3, the difficulty of Item 1 is 0.88 with a SEM of 0.11 for the male subgroup and 0.65 with a SEM of 0.10 for the female subgroup; the contrast in difficulty, 0.23, is the measure of DIF effect size (Linacre, 2010a); the Welch *t* value of this contrast is 1.52; and the *p* value of the contrast is .2101, which is not significant at the established threshold *p* value of .05.

UDIF analysis identified eight test items with significant DIF at $p < .05$: Items 6, 7, 21, 35, and 44 favoring male test takers, and Items 39, 43, and 49 favoring female test takers. Of these eight items, two (39 and 44) had UDIF magnitudes larger than 0.50 logits. The ICC of Items

TABLE 3
Uniform DIF Analysis of Items

Item	*Class A*	*DIF*	*DIF SE*	*Class B*	*DIF*	*DIF SE*	*DIF Contrast*	*df*	*Welch t*	*p*
1	Male	0.88	0.11	Female	0.65	0.10	0.23	842	1.52	.2101
2	Male	0.48	0.11	Female	0.65	0.10	−0.18	842	−1.18	.2693
3	Male	0.16	0.11	Female	−0.11	0.11	0.27	842	1.77	.1069
4	Male	−0.56	0.12	Female	−0.51	0.11	−0.05	842	−0.30	.7685
5	Male	−0.43	0.12	Female	−0.15	0.11	−0.27	842	−1.72	.0993
6	Male	0.72	0.11	Female	1.04	0.11	−0.31	842	−2.11	.0194
7	Male	0.01	0.11	Female	0.50	0.10	−0.49	842	−3.27	.0011
8	Male	−1.38	0.15	Female	−1.50	0.15	0.12	842	0.58	.4741
9	Male	−0.86	0.13	Female	−0.99	0.13	0.13	842	0.73	.6736
10	Male	−0.83	0.13	Female	−0.93	0.13	0.10	842	0.55	.4282
11	Male	−0.39	0.11	Female	−0.66	0.12	0.28	842	1.68	.0679
12	Male	0.74	0.11	Female	0.74	0.10	0.00	842	0.00	.8462
13	Male	0.35	0.11	Female	0.48	0.10	−0.13	842	−0.85	.5066
14	Male	0.42	0.11	Female	0.58	0.10	−0.16	842	−1.05	.3088
15	Male	−0.71	0.12	Female	−0.49	0.11	−0.22	842	−1.32	.1981
16	Male	0.41	0.11	Female	0.57	0.10	−0.16	842	−1.05	.3079
17	Male	0.74	0.11	Female	0.56	0.10	0.19	842	1.27	.1869
18	Male	0.31	0.11	Female	0.11	0.11	0.20	842	1.34	.1690
19	Male	0.07	0.11	Female	0.31	0.10	−0.24	842	−1.58	.1379
20	Male	0.15	0.11	Female	0.15	0.10	0.00	842	0.00	.8874
21	Male	0.28	0.11	Female	0.64	0.11	0.37	842	−2.46	.0284
22	Male	0.01	0.11	Female	0.20	0.12	−0.19	842	−1.24	.3678
23	Male	−0.10	0.11	Female	−0.17	0.11	0.07	842	0.45	.7202
24	Male	−0.88	0.13	Female	−0.81	0.11	−0.07	842	−0.41	.9589
25	Male	−0.18	0.11	Female	−0.24	0.10	0.06	842	0.35	.5966
26	Male	1.21	0.11	Female	1.08	0.11	0.13	842	0.85	.5526
27	Male	0.62	0.11	Female	0.62	0.10	0.00	842	0.00	.9452
28	Male	−0.25	0.11	Female	−0.08	0.11	−0.16	842	−1.03	.5544
29	Male	0.15	0.11	Female	0.30	0.10	−0.15	842	−0.97	.2972
30	Male	−0.18	0.11	Female	−0.34	0.11	0.15	842	0.97	.2374
31	Male	−0.03	0.11	Female	0.16	0.11	−0.19	842	−1.25	.1893
32	Male	−0.20	0.11	Female	−0.31	0.11	0.12	842	0.74	.4781
33	Male	−0.55	0.12	Female	−0.68	0.12	0.13	842	0.76	.5569
34	Male	−0.58	0.12	Female	−0.66	0.12	0.09	842	0.51	.4485
35	Male	−0.13	0.11	Female	0.22	0.10	−0.35	842	−2.32	.0278
36	Male	0.64	0.11	Female	0.52	0.10	0.12	842	0.82	.7899
37	Male	−0.25	0.11	Female	−0.10	0.11	−0.15	842	−0.96	.3158
38	Male	0.58	0.11	Female	0.35	0.10	0.23	842	1.53	.1674
39	Male	0.40	0.11	Female	−0.13	0.11	0.53	842	3.51	.0005
40	Male	0.47	0.11	Female	0.19	0.11	0.28	842	1.88	.0515
41	Male	0.22	0.11	Female	0.02	0.11	0.20	842	1.33	.2072
42	Male	−0.23	0.11	Female	−0.30	0.11	0.07	842	0.42	.6130
43	Male	−0.64	0.12	Female	−0.98	0.13	0.34	842	1.96	.0403
44	Male	0.35	0.11	Female	1.01	0.11	−0.66	842	−4.41	.0000
45	Male	−1.05	0.13	Female	−1.30	0.14	0.25	842	1.28	.1896
46	Male	−0.69	0.12	Female	−0.69	0.12	0.00	842	0.00	.6750
47	Male	−0.04	0.11	Female	0.04	0.11	−0.08	842	−0.51	.6656
48	Male	−0.17	0.11	Female	−0.17	0.11	0.00	842	0.00	.5209
49	Male	1.04	0.11	Female	0.56	0.10	0.48	842	3.23	.0042
50	Male	0.05	0.11	Female	−0.15	0.11	0.20	842	1.31	.1860

Note. Male = 425 participants; female = 427 participants. Items 1 to 15 belong to Section 1, Items 16 to 35 to Section 2, and Items 36 to 50 to Section 3. DIF = differential item functioning.



Note. 1 = Male (n = 425). 2 = Female (n = 427).

FIGURE 2 Item characteristic curves of Items 39 and 44 by gender (color figure available online).

39 and 44 are presented in Figure 2. The solid line in Figure 2 is the Rasch model curve. Test item 39 favors male test takers at ability levels up to –0.6 logits, and female test takers at ability levels between –0.6 and 2.1 logits (the domain containing most test takers). There is no difference above 3 logits, where few test takers landed. On test item 44, the gender subgroups' ICC curves intersect at three points: –1.2, 1.7, and 2.7 logits (horizontal axis). These are the turning points at which the subgroups' probabilities of correctly answering the item intersect. Inside a range of overall item performance from –2 to -1.2 logits, female test takers were more likely to answer this item correctly, whereas from –1.2 to 1.7 (the domain containing most test takers), male test takers were more likely to answer this item correctly; from 1.7 to 4, female test takers were more likely to answer this item correctly. At intersecting points in both items, both groups were equally likely to answer the item correctly. Because both test items favored either subgroup depending on ability level, both represent instances of NUDIF, indicating that DIF is likely a function of ability within the gender subgroups.

Figure 3 displays the results of UDIF analysis of all test items. The gray line drawn between square plot points represents local item difficulty in the female subgroup, and the black line drawn between diamond-shaped plot points represents local item difficulty in the male subgroup.

For most test items, gender difference in difficulty measures is small: As a whole, items appear not to display a great deal of gender-based UDIF. The two notable exceptions are Items 39 and 44.

Our DIF analysis as presented in Figure 2 appears to show an interaction between observed DIF and students' ability level in (at least) two items. We divided the gender subgroups into high- and low-ability subclasses by bisecting the range of person ability measures at the point in the middle of the range, and we performed a NUDIF analysis of all test items. Significant results of this analysis are presented in Table 4.

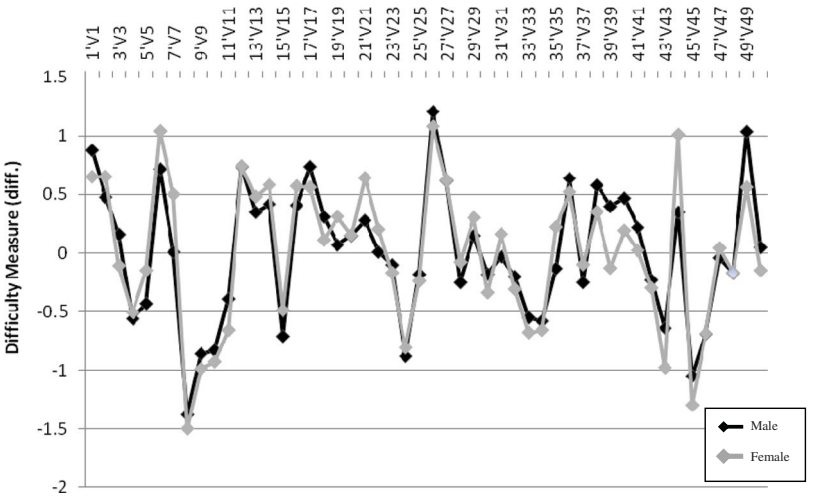


FIGURE 3 Uniform differential item functioning in Michigan English Language Assessment Battery listening test items.

TABLE 4
Nonuniform DIF Analysis of 50 Items

Item	*Class A*	*DIF Measure*	*DIFSE*	*Class B*	*DIF Measure*	*DIFSE*	*DIF Contrast*	*Welch t*	*p*
6	1.001	0.54	0.12	1.002	1.44	0.22	−0.90	−3.62	.0004
6	1.001	0.54	0.12	2.002	1.87	0.23	−1.33	−5.2	.0000
7	1.001	0.04	0.12	2.001	0.55	0.11	−0.51	−3.17	.0016
13	1.001	0.32	0.11	2.002	0.95	0.24	−0.63	−2.34	.0203
21	1.001	0.24	0.11	2.001	0.69	0.11	−0.45	−2.81	.0051
28	1.001	0.34	0.12	2.002	−0.29	0.28	0.63	−2.05	.0416
30	1.001	−0.07	0.12	1.002	−1.95	0.72	1.88	2.59	.0107
30	1.001	−0.07	0.12	2.002	−1.96	0.72	1.89	2.60	.0105
35	1.001	−0.04	0.12	1.002	−1.23	0.51	1.19	2.26	.0256
36	1.001	0.46	0.11	1.002	1.39	0.22	−0.93	−3.73	.0002
36	1.001	0.46	0.11	2.002	1.40	0.23	−0.94	−3.63	.0004
37	1.001	−0.17	0.12	2.002	−1.96	0.72	1.79	2.46	.0153
39	1.001	0.38	0.11	2.001	−0.13	0.11	0.52	3.19	.0015
41	1.001	0.12	0.12	1.002	0.68	0.25	−0.56	−2.03	.0440
41	1.001	0.12	0.12	2.002	0.77	0.25	−0.65	−2.33	.0211
43	1.001	−0.54	0.12	1.002	−2.65	1.01	2.11	2.08	.0396
43	1.001	−0.54	0.12	2.001	−0.91	0.13	0.37	2.03	.0424
43	1.001	−0.54	0.12	2.002	−2.67	1.01	2.12	2.09	.0388
44	1.001	0.29	0.11	2.001	0.92	0.12	−0.63	−3.83	.0001
44	1.001	0.29	0.11	2.002	1.40	0.23	−1.11	−4.3	.0000
49	1.001	0.99	0.12	2.001	0.55	0.11	0.44	2.68	.0076
50	1.001	0.04	0.12	2.002	−1.00	0.46	1.04	2.18	.0314

Note. 1.001 = lower ability male subgroup; 2.001 = lower ability female subgroup; 1.002 = higher ability male; 2.002 = higher ability female. DIF = differential item functioning.

When low-ability female and male participants are compared, 24 instances of significant NUDIF emerge. The significant NUDIF, as shown by their t values ($p < .001$), indicates that the magnitude of DIF varies according to the ability level of test takers. For example, NUDIF has occurred in test item 6 because it is easier for lower ability male subgroup (i.e., Class A in the first and second rows) who are likely to score higher on this item compared with higher ability male and female test takers (i.e., Class B in the first and second rows). This observation is counterintuitive. We discuss the possible reasons in details in the section next.

DISCUSSION

The unidimensionality, local independence, and reliability analyses of the test support Weigle's (2000) and Johnson's (2003) claims about the high reliability of MELAB test items and satisfy the preconditions for DIF analysis.

Determining the actual cause of observed DIF is often challenging (Camili & Shepard, 1994; Gierl, 2005). This problem prevails in exploratory DIF studies that lack an a priori hypothesis (Jang & Roussos, 2009). In this section, we explore the possible causes of DIF.

Test Section 1

Items 6, 7, and 13. Analysis of these test items reveals three common elements: (a) high difficulty, likely as a result of (b) long stems (see Uiterwijk & Vallen, 2005), and (c) NUDIF favoring low-ability male individuals. Item 6 is the second most difficult item on the test, and its stem is 17 words long, almost twice as long as many other items in this section. Its DIF contrast of 0.31 logits in favor of male test takers becomes more pronounced in NUDIF analysis (i.e., when ability levels are factored in), which reveals it to systematically favor low-ability male individuals over both high-ability male and female individuals. Similarly, Items 7 and 13 are relatively difficult (+0.26 and +0.42 logits, respectively), are long (16-word and 26-word stems), and favor low-ability male individuals over low-ability female individuals (by 0.51 and 0.63 logits).

It appears likely in all three cases that 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 (see Richardson, 1997a, 1997b).

Relatedly, there seems to be evidence that some low-ability test takers, especially male, were attracted to the distracters (incorrect answers) in some non-DIF items, such as Item 14, which is also among the most difficult items in Section 1 and has a lengthy stem. This item has two deliberately misleading distracters. Both contain two major key words repeated from the stem, when in fact the correct answer does not contain any of these words. It seems likely that low-ability test takers were attracted to these distracters and so did not attempt to guess, leading them to perform on the item as modeled (Linacre & Wright, 1994; Wright & Linacre, 1994). Respective infit and outfit statistics of 1.03 and 1.01 of Item 14 support this hypothesis.

This analysis is limited in that we did not have access to the data patterns of distracters in order to identify which distracters functioned satisfactorily or poorly. However, the binary scores on the DIF items in this section indicate that a number of lower ability male participants gave correct answers unexpectedly.

Test Section 2

Items 21, 28, 30, and 35. Our analysis identified substantive, significant UDIF in Items 21 and 35 and NUDIF in Items 28, 30, and 35.

On Item 28, high-ability female participants markedly outperformed low-ability male participants, but high-ability male test takers negligibly outperformed low-ability test takers in general. Our post hoc analysis suggests that this is probably because the item evaluates not only listening and inference-making ability but also test takers' short-term memories. In Item 28, a man and a woman have a short dialogue in which the woman complains about the transport system using several idiomatic expressions. The man asks a short question, and the woman gives a 39-word reply, the lengthiest in Section 2. The question requires test takers to rely on their schema in order to make an inference about the situation, and as they are not allowed to take notes in this section, they must also rely on their memory spans. High-ability test takers might be expected to function better on items tapping inference making skills (A. Wagner, 2004; E. Wagner, 2002) and the probable auxiliary dimensions of memory span (Dunkel et al., 1993), so the observed DIF can likely be ascribed to these considerations. It appears that the observed DIF in this item reflects a combination of ability and gender.

Items 30 and 35 were substantively more difficult for low-ability male test takers than for other test takers. To answer these items correctly, test takers must choose an option that accurately paraphrases the oral input. Previous research shows that high-ability listeners outperform lower ability listeners in this listening skill (A. Wagner, 2004), which resonates with our findings. In addition, the ultimate lines of these dialogues contain negative and double-negative phrases. Parsing and comprehending statements containing such phrases are probably more difficult for low-ability English learners (Cook, 2001) irrespective of their gender. Second-language learners usually learn negative morphemes after mastering simpler morphemes (Cook, 2001; Dimroth, 2010), and negative statements are an observed cause of DIF (Uiterwijk & Vallen, 1997, 2005). It seems that the DIF in these items pertains to ability level more than it does to gender, especially when the item is unlikely to contain any factor other than the negative morphemes.

This finding has implications for the proper use of Rasch model fit statistics, in particular the infit and outfit MNSQ indices. Item 30 tends to overfit the Rasch model, yielding "Guttman-like" (deterministic) results (Wright & Linacre, 1994, p. 370; i.e., many high-ability test takers and very few low-ability test takers answered it correctly). Under these circumstances, high- and low-ability people perform on the item as expected by the model (Linacre & Wright, 1994), but a 21% "deficiency in Rasch-model-predicted" randomness is still present (outfit = 0.79; 1–0.79 = 0.21; Wright & Linacre, 1994, p. 370), indicating that there exists [100 × (1–0.79)/0.79] 27% more defection in the observed performance than modeled: "The item difficulty estimated from low-ability persons differs noticeably from the item difficulty estimated from high-ability persons" (Wright & Linacre, 1994, p. 370). This pattern is also present but less marked in Item 35. This finding seems to show that establishing more stringent fit criteria in Rasch-based analysis

of dichotomous data can help detect erratic patterns and perturbation that might be due to a confounding effect in an item-level such as DIF (Smith, 1996). Wright and Linacre's (1994) range from 0.8 to 1.2 is therefore recommended.

Test Section 3

Items 39, 43, 44, and 49 were identified as significant UDIF cases, but only for Items 39 and 44 is this DIF substantive. Items 36, 37, 39, 41, 43, 44, 49, and 50 were identified as cases of significant NUDIF.

Items 36, 41, and 44. These items evaluate test takers' ability to paraphrase what they hear. UDIF is substantive only in 44, which favors male participants. Items 36 and 41 have significant NUDIF favoring low-ability male participants. These items tend to underfit the model, indicating some randomness in the response patterns of outlying test takers (in this case, low-ability male test takers). It is again possible that low-ability male participants attempted to guess the correct answer on these test items.

Items 37, 39, 43, 49, and 50. Items 43 and 49 show significant but not substantive UDIF. NUDIF in Item 49 is not substantive, but 43 shows substantive NUDIF favoring high-ability female and male participants over low-ability test takers. The low outfit MNSQ index of this paraphrase-type item (0.78) indicates high discrimination (Linacre & Wright, 1994; Pae & Park, 2006). That is, many high-ability test takers and very few low-ability test takers answered it correctly, forming Guttman response pattern for the item.

Items 37 and 50 are not UDIF cases but exhibit significant NUDIF. Item 39 slightly favors low-ability male test takers as opposed to low-ability female test takers. Items 37 and 50, which respectively evaluate test takers' ability to make paraphrases and identify details in an oral text about mummies, favor high-ability female and male test takers. These test items also tend to overfit the Rasch model.

In each of these overfitting test items, it appears likely that attractive distracters dissuaded low-ability male participants from making guesses and that the items' resulting high discrimination power may be the cause of DIF (Pae & Park, 2006).

CONCLUSION

This study set out to investigate the preconditions for Rasch-based DIF analysis (i.e., unidimensionality and local independence) in the MELAB listening test, gender-based UDIF and NUDIF and their causes, and whether the stringent Rasch fit criteria can indicate the presence of DIF.

The PCAR and Pearson correlations of residuals supported the preconditions for the Rasch model, so we investigated the other research objectives. Our study demonstrated a degree of gender-based DIF in the MELAB listening test, which became more evident when gender subgroups were split according to ability in NUDIF analysis. A possible explanation for the observed gender-ability NUDIF is that lower ability male test takers were more likely prone to take risks and displayed a pattern of successful lucky guesses on test items with unattractive distracters.

Test item content is also a likely cause of observed DIF. Some elements in DIF items may have been more difficult for low-ability test takers, including negative morphemes, idioms, metaphors, and items with lengths that may tax the memory spans of low-ability listeners (Vandergrift, 2004).

The linguistic factors previously described were handled better by high-ability test takers, who may have resorted to inferencing strategies by making use of linguistic or contextual cues where possible. These factors are by nature construct relevant and their DIF does not confound the interpretation and use of scores. However, the guessing factor and stem length are construct irrelevant and attenuate the validity argument of the test (see Aryadoust, in press). Test designers should limit the possibility of guessing correctly on test items by increasing answer options to four or even five effectively performing distracters. In addition, some distracters could be further improved to make them more attractive to test takers by linking them to possible prior knowledge or schema about a topic or to linguistic knowledge. This could also help oblige higher ability listeners to resort to using not only top-down (prior-knowledge-driven) but also bottom-up (text-driven) strategies to achieve their comprehension goals during the test, which might help to further differentiate performance within this group of test takers.

We also examined reliability and unidimensionality in the MELAB listening test. We found support for both, but the interactivity of decontextualized items is a matter of concern. These items may show high correlations with contextualized items, but their degree of utility cannot be established by correlations per se. We find Wright and Linacre's (1994) model fit criterion (0.8–1.20) more useful than other criteria in examining dimensionality and response patterns. Although PCAR showed that the test meets unidimensionality and local independence prerequisites to a considerable extent, examining fit statistics demonstrated that DIF items had erratic patterns. It seems that the presence of tiny percentages of noise (underfit) or deficiency (overfit) in data can point to serious problems, something that might not be detected by conventional dimensionality analyses such as PCAR.

We have alluded to the contribution of item format, the number of distracters and their attractiveness, and the cognitive features of test takers to DIF. These variables need to be studied more closely to better understand the factors that affect DIF in listening tests. We recommend that researchers conduct qualitative studies to investigate dimensionality, the incidence of guessing, and the selection of focal and reference groups and that they apply a confirmatory rather than an exploratory approach, seeking expert judgment and examining item writing guidelines to help formulate a hypothesis prior to undertaking DIF analysis. Finally, as Zumbo (2007) argued, DIF might be caused by contextual factors, which have not been examined closely in language assessment. Investigation of DIF caused by these variables—which marks the third generation of DIF studies—would be a promising line of inquiry in the field of language assessment.

ACKNOWLEDGMENTS

This study was made possible by a grant from the SPAAN Fellowship Program of the English Language Institute of the University of Michigan (ELI-UM) received by the first two authors. We thank ELI-UM for their supports and the reviewers of *LAQ*, Fred Meyer, Janna Fox, Fred Davidson, and Meg Malone for their comments on the article.

REFERENCES

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional Item Response Theory to evaluation educational and psychological tests. *Educational Measurement, Issues and Practice*, *22*, 37–50.

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. (2011). 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. (in press). Differential item functioning in while-listening performance tests. *The International Journal of Listening*.

Bond, T. G. (1994). Too many factors? *Rasch Measurement Transactions*, *8*, 347.

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

Boyle, J. (1987). Sex differences in listening vocabulary. *Language Learning*, *37*, 273–284.

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.

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

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Cauffman, E. (2006). A Rasch differential item functioning analysis of the Massachusetts Youth Screening Instrument identifying race and gender differential item functioning among juvenile offenders. *Educational and Psychological Measurement*, *66*, 502–521.

Coenen, M., & Vallen, T. (1991). Item bias in de Eindtoets Basisonderwijs [Item bias in the Final Test of Primary Education]. *Pedagogische Studiën*, *68*, 15–26.

Cook, V. (2001). *Second language learning and language teaching (3rd Edition)*. London, UK: Arnold/Oxford University Press.

Derwing, T. M., & Rossiter, M. J. (2003). The effects of pronunciation instruction on the accuracy, fluency and complexity of L2 accented speech. *Applied Language Learning*, *13*, 1–17.

Dimroth, C. (2010). The acquisition of negation. In L. R. Horn (Ed.), *The expression of negation* (pp. 39–73). Berlin, Germany: Mouton de Gruyter.

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.

Edelen, M. O., McCaffrey, D. F., Marshal, G. N., & Jaycox, L. H. (2009). Measurement of teen dating violence attitudes: An item response theory evaluation of differential item functioning according to gender. *The Journal of Interpersonal Violence*, *24*, 1243–1263.

Eom, M. (2008). Underlying factors of MELAB listening construct. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *6*, 77–94.

Ferber, M. A., Birnbaum, B. G., & Green, C. A. (1983). Gender differences in economic knowledge: A reevaluation of the evidence. *Journal of Economic Education*, *14*, 24–37.

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.

Flowerdew, J., & Miller, L. (2010). Listening in a second language. In A. D. Wolvin (Ed.), *Listening and human communication in the 21st century* (pp. 158–178). West Sussex, UK: Blackwell.

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.

Gierl, M. J. (2005). Using dimensionality-based DIF analyses to identify and interpret constructs that elicit group differences. *Educational Measurement: Issues and Practice*, *24*, 3–14.

Gierl, M. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *Journal of Educational Measurement*, *38*, 164–187.

Goh, C. C. M. (2005). Second language listening expertise. In K. Johnson (Ed.), *Expertise in second language learning and teaching* (pp. 64–84). Basingstoke, UK: 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, UK: Cambridge University Press.

Goh, C., & Aryadoust, S. V. (2010). Investigating the construct validity of MELAB listening test through the Rasch analysis and correlated uniqueness modeling. *Spaan Fellowship Working Papers in Second of Foreign Language Assessment*, *8*, 31–68.

Horn, L. R. (Ed.). (2010). *The expression of negation*. Berlin, Germany: Mouton de Gruyter.

Jang, E. E., & Roussos, L. (2007). An investigation into the dimensionality of TOEFL using conditional covariance-based nonparametric approach. *Journal of Educational Measurement*, *44*, 1–21.

Jang, E. E., & Roussos, L. (2009). Integrative analytic approach to detecting and interpreting L2 vocabulary DIF. *International Journal of Testing*, *9*, 238–259.

Johnson, J. (2003). *MELAB technical manual*. Ann Arbor: English Language Institute, the University of Michigan.

Klaassen, R. G., & Snippe, J. (1998). *Effective learning behaviour in English medium instruction: A pilot study*. Delft, the Netherlands: Delft University Press.

Kunnan, A. J. (1990). DIF in native language and gender groups in an ESL placement test. *TESOL Quarterly*, *24*, 741–746.

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. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *5*, 37–78.

Lin, J., & Wu, F. (2003, April). *Differential performance by gender in foreign language testing*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, Illinois.

Linacre, J. M. (1998a). Detecting multidimensionality: Which residual data-type works best? *Journal of Outcome Measurement*, *2*, 266–283.

Linacre, J. M. (1998b). Rasch first or factor first? *Rasch Measurement Transactions*, *11*, 603.

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

Linacre, J. M. (2010a). *A user's guide to WINSTEPS*. Chicago, IL: Winsteps.com.

Linacre, J. M. (2010b). WINSTEPS, Version 3.69 [Computer program]. Chicago, IL: Winsteps.com.

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

Lumsden, K. G., & Scott, A. (1987). The economics student reexamined: Male–female differences in comprehension. *The Journal of Economic Education*, *18*, 365–375.

Luppescu, S. (1993). DIF detection examined. *Rasch Measurement Transactions*, *7*, 285–6.

Maccartty, F. (2000). Lexical and grammatical knowledge in reading and listening comprehension. *Foreign Language Annals*, *34*, 439–445.

Mahoney, K. (2008). Linguistic influences on differential item functioning for second language learners on the national assessment of educational progress. *International Journal of Testing*, *8*, 14–33.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1994). Identification of non-uniform differential item functioning using a variation of the Mantel–Haenszel procedure. *Educational and Psychological Measurement*, *54*, 284–291.

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

Pae, T., & Park, G. P. (2006). Examining the relationship between differential item functioning and differential test functioning. *Language Testing*, *23*, 475–496.

Prieto Maranon, P., Barbero Garcia, M. I., & San Luis Costas, C. (1997). Identification of nonuniform differential item functioning: a comparison of Mantel–Haenszel and item response theory analysis procedures. *Educational and Psychological Measurement*, *57*, 559–569.

Richardson, J. T. E. (1997a). Conclusions from the study of gender differences in cognition. In P. A. Caplan, M. Crawford, J. S. Hyde, & J. T. E. Richardson (Eds.), *Gender differences in human cognition* (pp. 131–169). New York, NY: Oxford University Press.

Richardson, J. T. E. (1997b). Gender differences in cognition: Results from meta-analysis. In P. A. Caplan, M. Crawford, J. S. Hyde, & J. T. E. Richardson (Eds.), *Gender differences in human cognition* (pp. 3–29). New York, NY: Oxford University Press.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement*, *20*, 355–371.

Roussos, L. A., & Stout, W. F. (2004). Differential item functioning analysis. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 107–116). Thousand Oaks, CA: Sage.

Roznowski, M., & Reith, J. (1999). Examining the measurement quality of tests containing differentially functioning items: Do biased items result in poor measurement? *Educational and Psychological Measurement*, *59*, 248–269.

Ryan, K., & Bachman, L. (1992). Differential item functioning on two tests of EFL proficiency. *Language Testing*, *9*, 12–29.

Schumacker, R. E., & Linacre, J. M. (1996). Factor analysis and Rasch. *Rasch Measurement Transactions*, *9*, 470.

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

Swaminathan, H. (1994). Differential item functioning: A discussion. In D. Laveault, B. D. Zumbo, M. E. Gessaroli, & M. W. Boss (Eds.), *Modern theories of measurement: Problems and issues* (pp. 63–86). Ottawa, Ontario, Canada: University of Ottawa.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*, 361–370.

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

Uiterwijk, H. (1994). *De bruikbaarheid van de Eindtoets Basisonderwijs voor allochtone leerlingen* [The usability of the Primary Education Final Test for nonnative students]. Arnhem, the Netherlands: Cito, Instituut voor Toetsontwikkeling.

Uiterwijk, H., & Vallen, T. (1997). Onderzoek naar bias voor allochtone leerlingen in de Cito-Eindtoets Basisonderwijs [Research into bias for nonnative students in the Cito Primary Education Final Test]. *Pedagogische Studiën*, *74*, 21–32.

Uiterwijk, H., & Vallen, T. (2005). Linguistic sources of item bias for second generation immigrants in Dutch tests. *Language Testing*, *22*, 211–234.

Vandergrift, L. (2004). Listening to learn or learning to listen? *Annual Review of Applied Linguistics*, *24*, 3–25.

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. (2002). Video listening tests: A pilot study. *Working Papers in TESOL & Applied Linguistics, Teachers College, Columbia University*, *2*, 1–21.

Walstad, W. B., & Robson, D. (1997). Differential item functioning and male–female differences on multiple choice tests in economics. *Journal of Economic Education*, *28*, 155–171.

Weigle, S. C. (2000). Review of Michigan English Language Assessment Battery. *Language Testing*, *17*, 449–455.

Wright, B. D. (1994a). Comparing factor analysis and Rasch measurement. *Rasch Measurement Transactions*, *8*, 350.

Wright, B. D. (1994b). Local dependency, correlations, and principal components. *Rasch Measurement Transactions*, *10*, 509–511.

Wright, B. D. (1996a). Local dependency, correlations, and principal components. *Rasch Measurement Transactions*, *10*, 509–511.

Wright, B. D. (1996b). Reliability and separation. *Rasch Measurement Transactions*, *9*, 472.

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

Wright, B. D., & Stone, M. H. (1988). *Identification of item bias using Rasch measurement*. (Research Memorandum No. 55). Chicago, IL: MESA Press.

Wyse, A. E. & Mapuranga, R. (2009). Differential Item Functioning Analysis Using Rasch Item Information Functions. *International Journal of Testing*, *9*, 333–357.

Zeidner, M. (1986). Are English language aptitude tests biased towards culturally different minority groups? Some Israeli findings. *Language Testing*, *3*(1), 80–98.

Zeidner, M. (1987). A comparison of ethnic, sex and age bias in the predictive validity of English language aptitude tests: Some Israeli data. *Language Testing*, *4*(1), 55–71.

Zenisky, A., Hambleton, R., & Robin, F. (2003). Detection of differential item functioning in large scale state tests: A study evaluating a two-stage approach. *Educational and Psychological Measurement*, *63*, 51–64.

Zhang, Y., Matthews-Lopez, J., & Dorans, N. (2003, April). *Using DIF dissection to assess effects of item deletion due to DIF on the performance of SAT I: Reasoning sub-populations*. Paper presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, Chicago, IL.

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores. Ottawa, Ontario, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense. Retrieved from the British Colombia University Web site: http://educ.ubc.ca/faculty/zumbo/DIF/handbook.pdf

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

APPENDIX

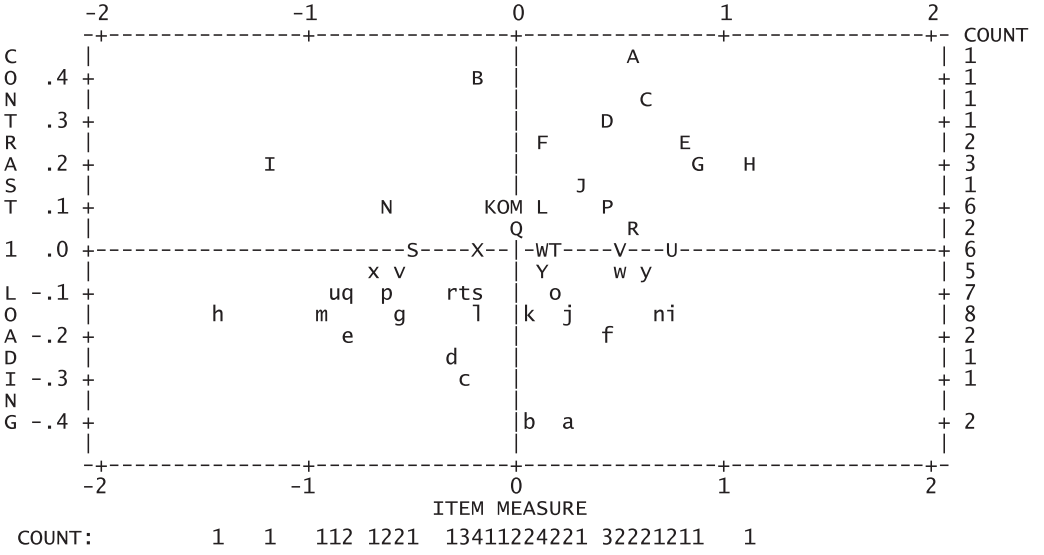


FIGURE A1 Illustration of the principal component analysis of linearized Rasch residuals of the data in the present study. *Note.* The vertical axis (contrast loading) represents the loading of item residuals on the first component extracted from the linearized Rasch residuals. The horizontal axis represents the item measure difficulty in logits. Each letter on the map represents an item.