International Journal of Testing, 15: 216–238, 2015

ISSN: 1530-5058 print / 1532-7574 online
DOI: 10.1080/15305058.2015.1004409

Routledge
Taylor & Francis Group

Fitting a Mixture Rasch Model to English as a Foreign Language Listening Tests: The Role of Cognitive and Background Variables in Explaining Latent Differential Item Functioning

Vahid Aryadoust
Centre for English Language Communication, National University of Singapore, Singapore

The present study uses a mixture Rasch model to examine latent differential item functioning in English as a foreign language listening tests. Participants ($n = 250$) took a listening and lexico-grammatical test and completed the metacognitive awareness listening questionnaire comprising problem solving (PS), planning and evaluation (PE), mental translation (MT), person knowledge (PK), and directed attention (DA). The listening test was subjected to MRM analysis where a two-latent class model had a sufficient fit. Next, an artificial neural network and a chi-square test were used to examine the nature of the latent classes. Class 1 comprised high-ability listeners capable of multitasking and obtained high PS, PE, and lexico-grammatical test scores but low DA, PK, and MT scores. Class 2 comprised low-ability listeners with limited multitasking skills who obtained high DA, PK, and MT scores but low scores on PS, PE, and the lexico-grammatical test. Finally, a model of listening comprehension is postulated and discussed.

Keywords: artificial neural network, gender, item response theory, lexico-grammatical knowledge, listening comprehension, metacognitive strategy awareness, mixture Rasch measurement

Listening comprehension is the ability to understand oral inputs, hold them in short-term memory, and build connections between them (Bodie & Crick, 2014).

Correspondence should be sent to Vahid Aryadoust, #2-24, Centre for English Language Communication, National University of Singapore, 10 Architecture Drive, 117511, Singapore. E-mail: vahid.aryadoust@nus.edu.sg

Listening is integral to language learning and communication in virtually every social context; nevertheless, it is considered the least researched language skill as indicated by the dearth of studies on second language (L2) listening test performance and differential learner attributes causing differential listening abilities (Vandergrift & Goh, 2012).

The primary focus of L2 listening research has been pedagogy (e.g., Janusik & Keaton, 2011), construct definition, and construct operationalization (e.g., Vandergrift & Goh, 2012). However, interrelations between listeners' cognitive and background attributes and their listening performance remain poorly understood even though they are regarded as key elements of listening constructs (Wolvin & Coakley, 1994). For example, Imhof and Janusik's (2006) model of listening identifies cognitive factors, such as lexico-grammatical knowledge and metacognition, and their interaction with listening contexts, but these factors remain critically underresearched.

Using a mixture Rasch model (MRM), the present study aims to investigate latent class differential item functioning (DIF) and its connections with test takers' cognitive and background factors. Latent class DIF occurs when test items function significantly differently across unobserved or latent groups within a certain population (Rijmen & De Boeck, 2005). The listening test used for this study is distinct in that it demands multitasking: test takers are required to simultaneously listen to the oral input, read the test items, and provide the best answers. To date, there has been little discussion concerning such listening tests (Aryadoust, 2012) and, to the best of my knowledge, this is the first study to undertake a latent class DIF analysis of these tests.

This study makes a contribution to the field of L2 listening comprehension assessment. In addition, the methods used to identify the latent classes and carry out post-hoc analysis have wider applications to all fields of educational assessment. By Drawing on Imhof and Janusik's (2006) model of L2 listening performance, the study explores a conceptual theoretical framework emerging from the available L2 listening literature. Rather than using students' raw scores, it identifies learners' latent classes on the basis of their listening test patterns and examines the interrelations between learners' class membership (i.e., groups of readers with highly similar test-taking patterns) and their attributes. As such, this study progresses toward a theory of differential L2 listening test performance where learners' cognitive and background attributes are used as the unifying basis for characterizing L2 listeners' group membership.

In terms of methodology, unlike previous research where manifest variables such as gender and nationality were used to identify differences in test performance (e.g., Aryadoust, 2012), this study uses MRM to detect the latent classes of L2 listeners. The study also applies an artificial neural network to investigate the relationship between the latent classes, metacognitive listening

strategy awareness, and lexico-grammatical knowledge. Finally, it uses a chi-square test of independence to examine the relationship between gender and the latent classes.

LITERATURE REVIEW

Listening Comprehension and Metacognitive Strategy Awareness

Although no single model of listening test performance can incorporate all possible complexities, Imhof and Janusik's (2006) listening model seems to have great potential for explaining listening comprehension mechanisms. Imhof and Janusik identified three primary factors affecting listening: (1) person-related attributes consisting of cognitive factors (e.g., lexico-grammatical knowledge, metacognition, and working memory) and affective factors (e.g., test anxiety and motivation). To these may be added demographic factors such as gender, age, and socio-economic variables to associate with learners' success in test performance (Dunkel, Henning, & Chaudron, 1993); (2) context-related attributes including formal or informal situations; and (3) results comprising quantitative outcomes (e.g., achieving comprehension) and qualitative outcomes (e.g., establishing relationships).

Successful listening results in the formation of a set of propositions or mental representations of audio input. The richness and accuracy of the propositions is intimately related to listeners' knowledge resources such as lexico-grammatical resources and world knowledge (Buck, 2001). A certain degree of comprehension is achieved when propositions are successfully generated and listeners can reconstruct the macrostructure of the audio input (Imhof & Janusik, 2006; Janusik, 2007).

In addition, metacognitive listening strategies play a significant role in listening comprehension (Goh, 2000). Through metacognition, learners become self-aware of their cognitive strategy use and can actively monitor and adjust these strategies to fulfill certain goals (Flavell, Miller, & Miller, 1993). Metacognition entails task knowledge (i.e., learners' grasp of the prerequisites of learning tasks and the factors affecting task difficulty), self-awareness (i.e., learners' awareness of their own anxiety, self-confidence, and reactions to learning prerequisites), and strategy use (learners' knowledge of learning techniques to employ in order to meet their learning objectives) (Goh & Hu, 2014).

Metacognitive strategies in listening comprehension have been conceptualized as a multidimensional construct by Vandergrift, Goh, Mareschal, and Tafaghodtari (2006) who developed a metacognitive awareness listening questionnaire (MALQ) to assess them. MALQ consists of 21 survey items engaging five primary dimensions: directed attention, mental translation, planning and evaluation, problem solving, and person knowledge. More recently, Vandergrift and Goh

(2012) grouped the first four factors (directed attention, mental translation, planning and evaluation, and problem solving) as a collective representation of learners' attempts to regulate the comprehension process and described the last factor (person knowledge) as a representation of learners' self-awareness. In all, it seems that lexico-grammatical resources, metacognitive listening strategy awareness, and listening have interconnections. The present study seeks to investigate this interdependence.

Differential Item Functioning

Overview. DIF occurs when test takers with the same ability from different backgrounds persistently have different probabilities of answering (a set of) test items accurately. Traditional DIF studies show that test scores are affected by a secondary and unmodeled dimension or construct besides the primary dimension (Aryadoust, Goh, & Lee, 2011).

Secondary dimensions may be either auxiliary (simple)—intended to be operationalized (e.g., vocabulary knowledge in a reading test)—or nuisance (complex)—intrusive to learners' test performance and contaminative to their scores (e.g., reading proficiency in a mathematics test). DIF caused by auxiliary dimensions is called benign, while DIF caused by nuisance dimensions is called adverse (Ackerman, Gierl, & Walker, 2003).

Multiple DIF detection techniques are currently available, including Rasch model DIF analysis, Mantel-Haenszel, and multidimensional item response theory (Fukuhara & Kamata, 2011). These techniques commonly use manifest variables (e.g., nationality, age, and gender) for characterizing DIF. Nevertheless, studies show that manifest variables might be inappropriate for explaining test takers' differential cognitive processes that result in DIF (Kubinger, 2005) because these variables are typically chosen according to analysts' speculations rather than knowledge of test takers' cognitive processes (Cohen & Bolt, 2005). Research has further revealed that subgroups based on manifest variables might be "heterogeneous" due to unmodeled variables that remain unnoticed (Maij-de Meij, Kelderman, & van der Flier, 2008).

Investigating Differential Item Functioning Through Mixture Rasch Model. This study adopts MRM, a psychometric technique that integrates Rasch measurement and latent class analysis, and computes the item difficulty parameters for each identified latent class individually. The issue of "heterogeneity" of subgroups of test takers can be resolved by MRM where the identified latent classes are characterized by their conditional probabilities or the odds that variables take on certain magnitudes (Rost, 1990).

There are two methods of MRM analysis: a method where covariates can be examined at the same time as the estimation of the latent classes (Dai, 2013;

Smit, Kelderman, & van der Flier, 2000) and a two-stage analysis method where no covariate is used. MRM with covariate is a confirmatory approach in which there is a priori substantive evidence that the covariate mediates person and item parameters (Lin & Lin, 2013). According to Lin and Lin (2013), "The more distinct the latent classes [are], the easier the MRM model with a covariate [can] separate examinees into different latent groups" (p. 393). Nevertheless, most MRM analyses, due to the lack of such information, adopt the two-stage MRM approach.

In the first stage of the two-stage approach, several competing models are estimated and the most appropriate one is chosen based on the fit of the model (Rost & von Davier, 1994). Three fit statistics have been suggested for assessing fit in MRM analysis: Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Consistent Akaike Information Criterion (CAIC). AIC, however, has been shown to be less accurate because it is sensitive to sample size (Baghaei & Carstensen, 2013; Li, Cohen, Kim, & Cho, 2009). The most appropriate model will have the smallest fit statistics.

In the second stage of the two-stage approach, learners' cognitive strategies and background information may be used to explore the distinctively qualitative features of each emerging latent class (Cohen & Bolt, 2005). For example, Hong and Min (2007) detected three latent classes where gender had a significant but weak impact, while Rost, Carstensen, and von Davier (1997) identified extreme and moderate students in a study of personality assessment. In a more recent study, Baghaei and Carstensen (2013) identified two latent classes in a reading test: one class consisting of learners with high scores on short-context test items and the other comprising high-ability learners with higher scores on long-context test items.

Nevertheless, there has been little empirical research on DIF in listening tests. Two recent studies by Aryadoust (2012) and Aryadoust and colleagues (2011) addressed DIF across gender, prior exposure to similar tests (practice effect or test-wiseness), nationality, and age. Both studies reported the presence of gender-based DIF, likely caused by male test takers' tendency for lucky guessing, which is rewarded by a decent chance (33%) of getting the right answer among three options. Prior exposure to tests also induces DIF, whereas nationality and age do not induce DIF.

Aryadoust's (2012) study has several commonalities with the present study; notably, both use a listening test where multitasking seems to be an important mechanism required for answering the test items. Learners taking these tests should listen to the oral input while reading test items and choosing or supplying the most appropriate responses. As Field (2009) argued, these (virtually) simultaneous cognitive processes will tax listeners' working memory and cognitive resources, putting less-experienced test takers at a disadvantage and likely inducing DIF. Many well-known high-stakes tests of listening, including the International English

Language Testing System (IELTS) and the Business English Certificate (BEC) Vintage, seem to engage test takers' ability to multitask (Aryadoust, 2012).

METHODOLOGY

Participants

Two-hundred and fifty first- and second-year English as a foreign language (EFL) college students aged between 17 and 23 years ($M = 19.73$; $SD = 0.90$) from the People's Republic of China consented to participate in the study and were administered multiple psychometric instruments. After participation, a personalized test performance report was generated for each student. Both the assessment and report were partly designed to aid participants in their preparation for English exams, which contain lexico-grammatical and listening sections.

In previous studies, the sample size for MRM analysis has varied between 250 and 5500. For example, Frick, Strobl, and Zeileis's (in press) study contained 273 subjects (class 1 = 111; class 2 = 53; class 3 = 109), whereas Chen and Jiao (2014) had access to data from 5233 subjects who participated in the Program for International Student Assessment (PISA) project (class 1 = 1988; class 2 = 3245). While psychometric modeling with large data sets is highly desirable, in practice, readily available data such as PISA are extremely "sparse" (Chen & Jiao, 2014). In addition, there is as yet no freely available listening comprehension data set for researchers who wish to explore listening—the most underresearched language skill (Aryadoust, 2012). Therefore, listening comprehension researchers collecting data on their own face the logistic and financial constraints of data collection. Another important challenge is motivating students and encouraging them to make their best efforts when taking the test so as to prevent missing and sparse data. The present study offered to provide feedback on lexico-grammatical knowledge and listening ability to the participating students. The students benefit from such feedback because the listening test closely resembles the BEC listening subtest they are required to take later in their studies.

Instruments

Data were collected from participants' performance on a listening test, a lexico-grammatical knowledge test consisting of vocabulary and grammar subsections, and the MALQ.

The Listening Test. A sample listening paper under the BEC Vintage intermediate level, a certification widely acquired by Chinese students seeking employment in industries calling for English language proficiency, was used for the listening test in this study. The test consisted of 30 test items of which items 1 to

TABLE 1
Sentence and Word Count, Sentence Length, Type-Token Ratio, and Flesch-Kincaid Grade Level Indices of the Three Listening Test Sections

Feature	Section 1	Section 2	Section 3
Word count	182.66	394.50	675.00
Sentence count	20.66	30.00	45.00
Sentence length	9.29	13.14	15.00
Flesch-Kincaid grade level	4.73	5.34	7.69

12 were fill-in-the-gap items based on phone conversations or phone messages, items 13 to 22 were fill-in-the-gap items based on five short recordings depicting a problem situation, and items 23 to 30 were multiple choice (MC) items based on a recorded interview with a restaurant manager. Participants were given time to read the questions prior to listening to the texts and each audio recording was played twice as per standard BEC listening test procedure.

Table 1 displays the descriptive statistics of the listening test. Word and sentence counts are the mean scores across the listening texts. The Flesch-Kincaid Grade level index quantifies the comprehension difficulty of the texts. All indices increase as we move through the sections, that is, section 3 has the highest word and sentence count and length as well as Flesch-Kincaid Grade level. It is therefore plausible to presume that as we move through the test sections, the text difficulty, cognitive demands, and working memory overload will increase.

As previously stated, the BEC demands simultaneous listening, reading, and answering of items. This format has two important consequences. First, learners have to shift their attention constantly between oral and written modalities, likely causing confusion and introducing construct-irrelevant variance (Aryadoust, 2012). Second, due to the demand for multitasking—which seems to be irrelevant to listening comprehension skills (Field, 2009)—strategic students who have received prior test-taking training are advantaged by the test structure.

The Lexico-Grammatical Test. The lexico-grammatical test comprises two sections: a vocabulary knowledge subtest and a grammatical knowledge subtest. The vocabulary knowledge subtest consisted of 30 MC items ranging in difficulty level to discriminate among participants. For each item, participants are instructed to select the synonym for an underlined target word among the provided options. The grammatical knowledge subtest consisted of 15 MC items selected from a sample paper-based test of English as a foreign language. The items measured participants' ability to recognize a range of grammatical structures, such as independent clauses, infinitives, and prepositions. Students were given their vocabulary and knowledge test results as diagnostic feedback.

The Metacognitive Awareness Listening Questionnaire. The MALQ includes five subscales: directed attention (four items), mental translation (three items), person knowledge (four items), planning and evaluation (five items), and problem solving (five items) and adopts a six-point Likert scale ranging from *strongly disagree* (1) to *strongly agree* (6).

DATA ANALYSIS

I used four primary data analysis methods in the present study. First, I examined the unidimensionality of the lexico-grammatical test and the MALQ components using the principal component analysis of linearized residuals (PCAR), which compares the observed variance to the "variance components expected for these data if they exactly fit the Rasch model" (Linacre, 2014, p. 280). In unidimensional tests, the components (contrasts) extracted by PCAR from the residuals explain merely a negligible amount of variance (approximately two eigenvalues), with the Rasch model dimension explaining a significantly larger amount of variance (Linacre, 2014).

Next, I subjected the listening test to MRM analysis. Students' performances on the MALQ and lexico-grammatical knowledge tests were separately calibrated using Rasch measurement (Andrich, 1978). Finally, an artificial neural network was used to classify the listening latent classes in terms of learners' MALQ and lexico-grammatical knowledge. To assess whether gender would differentiate the latent classes, a chi-square test of independence was performed.

Preliminary Item Response Theory Modeling

A recent study conducted by Alexeev, Templin, and Cohen (2011) suggested that if a test conforms to, for example, the two or three parameter logistic (2PL or 3PL) Item Response Theory (IRT) models, applying a MRM may result in an overextraction of latent classes. It is therefore important to initially explore the conformity of the tests to 2PL and 3PL IRT models, unless there is a valid reason why the models would not provide as much useful information as MRM. In the present study, the listening comprehension test does not lend itself to the 3PL model due to its multi-item format structure, which comprises fill-in-the-gap and multiple choice questions. Accordingly, a 2PL IRT model was fitted to the data and compared with the MRMs.

Mixture Rasch Model Analysis of the Listening Test

The BEC listening test was subjected to MRM analysis using the *WINMIRA* 32 computer package (von Davier, 2001a). Four models comprising one to four latent

classes were estimated and their fit indices were compared to choose the most parsimonious one.

Item difficulty, measurement error, and fit indices were computed for each latent class. Item difficulty is expressed in logits (log-odd-units) and measurement error statistics represent the precision item endorsibility statistics. For instance, a measurement error coefficient of 0.15 for an item of 2.00 logits difficulty indicates that the actual item difficulty index falls between 1.90 and 2.10 (von Davier, 2001a).

I examined the fit of the items using the *Q*-index and *ZQ*-index (Rost & von Davier, 1994). *Q*-index values have a range between zero and unity, with values closer to zero being desirable and values larger than 0.50 suggesting that the data is likely confounded by a secondary construct (von Davier, 2001b). *ZQ*-index is the *z*-standardized form of *Q*-index and ranges between $-\infty$ and $+\infty$, with indices falling outside of –1.96 and +1.96 suggesting misfit (von Davier, 2001a). The statistical significance of the observed misfit is indicated by its *p*-value. Finally, the degree of DIF per item was calculated using the χ^2 formula, which is expressed as $\chi^2 = \beta^2_{\text{diff}} / [v(\beta_1)+v(\beta_2)]$, where β^2_{diff} is the difference in item difficulty estimates across the latent classes and $v(\beta_1)$ and $v(\beta_2)$ are the variance of the item difficulty estimates for classes 1 and 2, respectively (Chen & Jiao, 2014; Oliveri, Ercikan, & Zumbo, 2013; Samuelsen, 2005).

Rasch Model Analysis of the Lexico-Grammatical Test and Metacognitive Awareness Listening Questionnaire

I examined the psychometric validity of the lexico-grammatical test and MALQ by estimating the item and person parameters, measurement error, and fit statistics. There are two main types of fit statistics in Rasch measurement: outfit and infit mean square (MNSQ) and *z*-standardized values. Outfit MNSQ is an index based on chi-square statistic that is sensitive to erratic patterns in outlying data points (i.e., items distant from test takers' ability). MNSQ itself is the chi-square statistic divided by its degree of freedom. In contrast, infit MNSQ statistics are sensitive to erratic patterns in inliers or items with difficulty levels close to test takers' ability and, accordingly, can capture erratic response patterns in items near test takers' ability. *Z*-standardized values are analogous to the *ZQ*-index in MRM and are interpreted in the same way.

Distinguishing Latent Classes Through Artificial Neural Network

To determine the differential features of the latent classes, the present study uses an artificial neural network analysis, which offers several advantages over the regression models and *t*-tests used in previous research (Baghaei & Carstensen, 2013). Unlike *t*-tests and regression models, artificial neural network does not

assume linearity of relationships between variables or a normal distribution and has a flexible structure that can emulate data (i.e., predict and classify) with high accuracy. In addition, it randomly splits the data into training and testing subsamples; the training subsample is used for detecting the (mathematical) relationships between dependent and independent variables and the testing subsample is used for verifying the relationship.

In this study, the latent classes emerging from MRM were set as the dependent variable and the learners' lexico-grammatical ability and metacognitive listening strategy awareness measured by MALQ were the predictors. Because the dependent variable was categorical (with two levels: Class 1 and 2), the artificial neural network took the role of a *classifier* to predict the dependent variable from the independent variables. The classification accuracy was assessed by estimating the proportion of accurately classified learners.

Finally, to examine the distribution of gender across latent classes, a chi-square test of independence was carried out.

RESULTS

Unidimensionality of the Tests

The MALQ subscales and the lexico-grammatical test were individually subjected to PCAR. These analyses yielded small variance estimates for residuals (eigenvalues < 2) and large variance estimates for Rasch model dimension, supporting the unidimensionality of the instruments.

Mixture Rasch Model Analysis of the Listening Test

Determining the Latent Classes. Initially, the 2PL IRT model was fitted to the data, but as its AIC (6667.39) and BIC (6862.37) indices indicate, it had a poor fit relative to the MRMs; this provides evidence that the 2PL item response functions cannot precisely emulate the constituent structure of the data. Next, four MRMs were estimated and compared to determine the number of latent classes underlying the listening test. Table 2 shows that the one-, three-, and four-latent class models have larger AIC, BIC, and CAIC statistics than the two-latent class model. Therefore, the two-latent class model was chosen over the other models as it has the lowest AIC (5154.54), BIC (5377.14), and CAIC (5440.14) statistics.

Table 3 displays class-specific information concerning the two-latent class model: the left column demonstrates the latent class, columns two and three give the expected and observed class size, and columns four and five provide the mean probability of class assignments. The latent classes were expected to comprise approximately 68% and 30% of the sample (class assignment for 2% of

TABLE 2
Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Consistent Akaike Information Criterion (CAIC) Coefficients of the Two Parameter Logistic (2PL) Model and the Four Mixture Rasch Models

Latent Class	AIC	BIC	CAIC
2PL Item Response Theory model	6667.39	6862.37	NA
One-latent class	5171.82	5381.36	5512.36
Two-latent class	**5154.54**	**5377.14**	**5440.14**
Three-latent class	5160.05	5495.72	5590.72
Four-latent class	5159.22	5607.96	5734.96

Note. Bold values indicate the optimal model.

the sample is unknown), which approximates the observed size (63% and 37%), indicating conformity between the model-expected and observed results. The mean probability columns demonstrate the average likelihood for participants classified in both classes. It is evident from the table that the off-diagonal statistics are much smaller than the diagonal indices, suggesting high classification precision (Hong & Min, 2007). For example, learners with a probability of 0.094 of being in latent Class 1 had a much higher likelihood (0.906) of being classified in latent Class 2.

Item Difficulty and Fit Statistics Across the Latent Classes. The class-specific item difficulty statistics estimated in the two-latent class model are graphically displayed in Figure 1: Class 1 is represented by a dotted line and Class 2 by a solid line. The item number on the horizontal axis is plotted against the item difficulty on the vertical axis. Item 3 (difficulty = 6.90 logits) is the most difficult item for both classes, but items 4 and 5, which are also difficult for Class 1 (difficulty = 6.90 logits), proved to be easy for Class 2 (difficulty = –2.00 and –0.10 logits). The easiest item for Classes 1 and 2 are items 30 and 23 (difficulty = –4.70 and –0.10 logits). In addition, 17 out of 30 items favor Class 1 and, of these, items 6 and 9 are in Part 1 and the remainder are in Parts 2 and 3. The rest of the items favor Class 1.

TABLE 3
Class-Specific Information Concerning the Two-Latent Class Model

Class	Expected Size	Observed Size	Mean Probability Class 1	Mean Probability Class 2
1	0.677 (68%)	0.626 (63%)	0.884	0.116
2	0.300 (30%)	0.373 (37%)	0.094	0.906

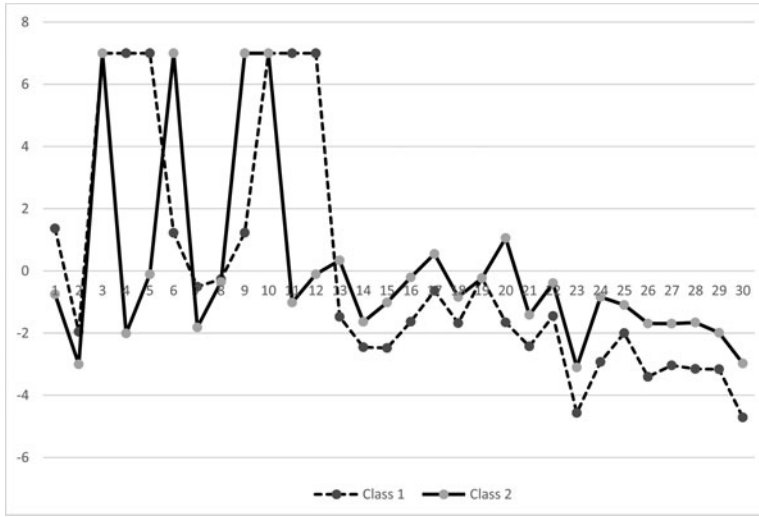

FIGURE 1
Item difficulty across latent Classes 1 and 2. *Note*. 17 out of 30 items favor Class 1, including 6, 8, 9, 14, 15, 18, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, and 30. In contrast, Items 1, 2, 4, 5, 7, 11, and 12 from Part 1 favor Class 2.

The differences between the difficulty parameters of each item across the two classes were calculated and converted to absolute numbers. The mean score of the absolute numbers is 2.38, indicating a significant difference between Classes 1 and 2 in favor of Class 1. To examine the differences in test taker ability across latent classes, I performed a *t*-test comparing the mean score of Class 1 (0.31) and Class 2 (0.10). The results showed a significant effect for latent class, $t(248) = 4.89$, $p < 0.001$, with Class 1 receiving significantly higher scores than Class 2.

Table 4 provides the *Q*-indices, *ZQ*- indices, their *p*-values, and the degree of DIF. The *Q*-indices for all items except item 2 in Class 2 are less than 0.50, suggesting that, on average, low-ability learners had a lower chance of answering the items accurately, whereas high-ability learners had a higher likelihood. The sufficient item fit indicates high discrimination power of the test items. The two far-right columns of Table 4 give the χ^2 values or differences in the item difficulty parameters across the two latent classes along with their associated labels including negligible ($p > 0.05$), moderate ($p < 0.05$ and degree of DIF < 10), and large ($p < 0.05$ and degree of DIF ≥ 10). Only 5 out of 30 items do not function differentially across the two classes (i.e., 3, 10, 13, 16, and 17). Items 3 and 10 had a very low variance (nearly zero), so the χ^2 index was not calculated for them.

TABLE 4
Fit Statistics, *p*-Value, and Degree of Differential Item Functioning

	Latent Class 1			Latent Class 2				
Item	*Q*-index	*ZQ*-index	*p*-value	*Q*-index	*ZQ*-index	*p*-value	Degree of DIF	Associated Label
L_1	0.302	0.060	0.475	0.181	−0.078	0.531	16.100*	Large
L_2	0.109	−0.743	0.771	0.413	2.002	0.022*	9.049*	Moderate
L_3	0.490	−0.055	0.522	0.490	−0.070	0.527	NA	Negligible
L_4	0.490	−0.055	0.522	0.385	1.325	0.092	79.707*	Large
L_5	0.490	−0.055	0.522	0.425	0.712	0.238	232.354*	Large
L_6	0.363	0.009	0.496	0.490	−0.070	0.527	935.884*	Large
L_7	0.127	−0.969	0.833	0.347	1.432	0.076	9.609*	Moderate
L_8	0.163	−0.444	0.671	0.315	0.498	0.308	7.738*	Moderate
L_9	0.247	−0.150	0.559	0.490	−0.070	0.527	935.848*	Large
L_10	0.490	−0.055	0.522	0.490	−0.070	0.527	NA	Negligible
L_11	0.490	−0.055	0.522	0.237	0.289	0.386	121.373*	Large
L_12	0.490	−0.055	0.522	0.138	−0.339	0.632	232.354*	Large
L_13	0.173	−0.133	0.553	0.348	0.004	0.498	0.371	Negligible
L_14	0.250	0.623	0.266	0.114	−0.960	0.831	5.485*	Moderate
L_15	0.138	−0.311	0.622	0.387	0.291	0.385	4.130*	Moderate
L_16	0.145	−0.250	0.598	0.246	−0.169	0.567	2.972	Negligible
L_17	0.168	0.024	0.490	0.247	−0.398	0.654	1.699	Negligible
L_18	0.187	0.017	0.493	0.262	0.010	0.495	5.398*	Moderate
L_19	0.233	0.276	0.391	0.376	0.219	0.413	7.099*	Moderate
L_20	0.185	−0.150	0.559	0.110	−0.574	0.717	6.776*	Moderate
L_21	0.155	0.038	0.484	0.245	−0.406	0.657	5.084*	Moderate
L_22	0.192	0.044	0.482	0.253	−0.198	0.578	4.379*	Moderate
L_23	0.243	0.132	0.447	0.170	−0.532	0.702	6.842*	Moderate
L_24	0.172	−0.117	0.546	0.279	−0.555	0.710	3.322*	Moderate
L_25	0.180	0.089	0.464	0.419	0.488	0.312	5.302*	Moderate
L_26	0.185	0.040	0.483	0.133	−1.061	0.855	4.406*	Moderate
L_27	0.179	0.169	0.432	0.205	−0.782	0.783	4.646*	Moderate
L_28	0.166	0.122	0.451	0.213	−0.861	0.805	4.547*	Moderate
L_29	0.251	0.359	0.359	0.306	0.062	0.475	5.019*	Moderate
L_30	0.186	0.161	0.436	0.214	−0.998	0.840	6.586*	Moderate

*$p < 0.05$.

Rasch Measurement of Metacognitive Awareness Listening Questionnaire Subscales and Lexico-Grammatical Test

The Rasch model reliability and fit were assessed individually for the lexico-grammatical test and the five subscales of MALQ. Table 5 displays the item reliability, separation coefficients, and infit and outfit MNSQ statistics for items and persons. Overall, all instruments achieved high reliability, indicating sufficient

TABLE 5
Reliability, Separation, and Fit Statistics of the Metacognitive Awareness Listening Questionnaire Subscales and Lexico-Grammatical Test

Subscale	Item Reliability	Item Separation	Average Person Infit MNSQ	Average Person Outfit MNSQ	Average Item Infit MNSQ	Average Item Outfit MNSQ
Directed attention	0.98	7.63	0.99	1.05	1.00	1.05
Mental translation	0.98	7.72	0.99	1.02	1.01	0.98
Person knowledge	0.98	7.62	1.07	0.96	1.06	0.98
Problem solving	0.88	2.66	1.00	1.00	1.01	1.01
Planning & evaluation	0.88	2.66	1.01	1.01	1.02	1.02
Lexico-grammatical test	0.98	6.55	1.00	1.01	1.00	1.01

Note. MNSQ, mean square.

discrimination between easy and difficult items. Planning and evaluation and problem solving have the lowest reliability and separation coefficients of 0.88 and 2.66, respectively, suggesting approximately three statistically distinct levels of item endorsibility. Directed attention, mental translation, and person knowledge have the highest reliability and separation coefficients of 0.98 and 7.63, 7.72, and 7.62, respectively, indicating approximately seven distinct levels of item endorsibility.

Furthermore, item and student have a sufficient fit to the model with the average infit and outfit MNSQ indices ranging between 0.96 and 1.06, indicating their conformity with the expectations of Rasch measurement and the lack of construct-irrelevant factors. Similarly, the lexico-grammatical test fits the model sufficiently with approximately seven levels of item difficulty (separation = 6.55; reliability = 0.98). Overall, these results support the psychometric validity of the instruments.

Artificial Neural Network Analysis and Chi-Square Test

Figure 2 illustrates the mean scores of the MALQ subscales and lexico-grammatical test for Classes 1 and 2. Class 1 has a higher average score on problem solving ($0.76 > 0.65$), planning and evaluation ($0.17 > -0.01$), and the lexico-grammatical test ($0.91 > -2.35$), whereas Class 2 has a higher average score on directed attention ($0.22 > 0.07$), person knowledge ($0.50 > 0.47$), and mental translation ($0.36 > 0.20$).

The artificial neural network analysis used 168 (66.4%) pieces of the data for training the algorithm and mapping the mathematical relationship between the independent and dependent variables. It then tested the model yielded during training across 85 (33.6%) pieces of the left-out data (two cases were excluded by the model, likely due to fit problems). Table 6 shows the prediction accuracy

TABLE 6
Prediction Accuracy of the Artificial Neural Network Across Classes 1 and 2

Sample	Observed	Predicted		
		Class 1	Class 2	Percent Correct
Training	Class 1	134	3	97.80%
	Class 2	23	8	25.80%
	Overall percent	93.50%	6.50%	84.50%
Testing	Class 1	72	3	96.00%
	Class 2	6	4	40.00%
	Overall percent	91.80%	8.20%	89.40%

information: the accuracy of prediction for Class 1 was significantly higher than the prediction accuracy for Class 2 across both training and testing subsamples (Class 1: 97.8% and 96.0%; Class 2: 25.8% and 40.0%). Overall, the correct prediction percentages for the training and testing subsamples were 84.5% and 89.4%, which are significantly high.

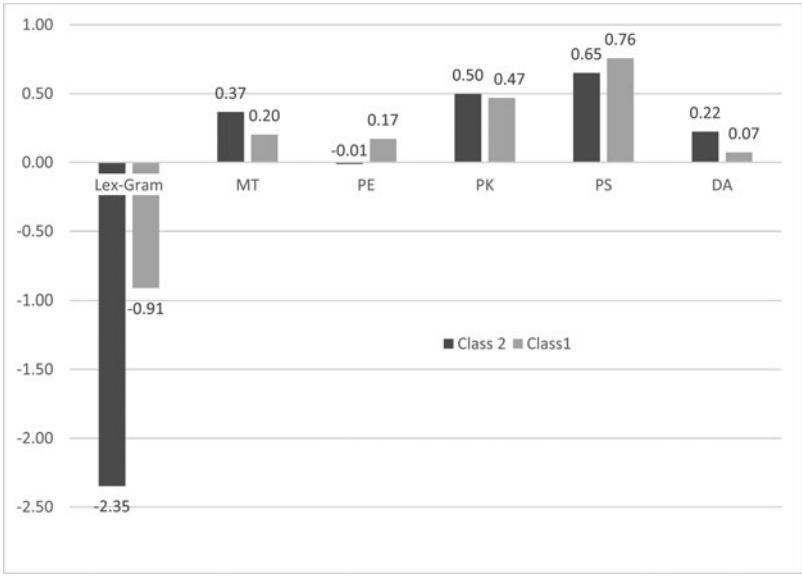

FIGURE 2
Illustration of the mean scores of the Metacognitive Awareness Listening Questionnaire (MALQ) subscales and lexico-grammatical test across the two latent classes. *Note*. DA, directed attention; Lex-Gram, lexico-grammatical knowledge; MT, mental translation; PE, planning and evaluation; PK, person knowledge; PS, problem solving.

TABLE 7
Mean Statistics of the Independence Variables Across the Latent Traits and Their Importance Values

Independent Variable	Mean Score for Class 1	Mean Score for Class 2	Variable Importance
Lexico-grammatical test	−0.911	−2.347	100.0%
Problem solving	0.757	0.650	64.9%
Person knowledge	0.468	0.497	57.7%
Mental translation	0.202	0.366	55.5%
Planning & evaluation	0.171	−0.011	42.8%
Directed attention	0.073	0.223	24.5%

Next, the variable importance (VI) statistics were estimated. VI shows the strength of association between the inputs or the independent variables and the output or the dependent variable (Aryadoust & Goh, 2014). As Table 7 shows, all independent variables have a role in determining the class membership of the learners. The most important variable is lexico-grammatical knowledge (VI = 100%), followed by problem solving (VI = 64.9%), person knowledge (VI = 57.7%), mental translation (VI = 55.5%), planning and evaluation (VI = 42.8%), and directed attention (VI = 24.5%). Finally, the chi-square test of independence showed no relationship between gender and class membership.

DISCUSSION

The present study was designed to determine the response patterns that form latent classes of EFL listeners and investigate the relationship between latent class membership, lexico-grammatical knowledge, metacognitive listening strategy awareness, and gender, using MRM, the artificial neural network, and chi-square tests.

Listening Latent Class Membership

Two latent classes with distinct response patterns emerged from the MRM analysis. Analysis of Class 1, which had a relatively higher listening ability, reveals that learners in this class were favored by 19 out of 30 test items, most of which belong to Parts 2 and 3. Items in Part 2 seem to engage listeners in multitasking: listeners are required to listen to the text while reading a list of eight activities of which the five activities described by the speaker must be chosen. To choose the accurate options, listeners must write the letters (A to H) corresponding to the options next to the number of the test item. This set of simultaneous cognitive-motor tasks does not seem to directly relate to listening comprehension (Dunkel et al., 1993; Field,

2009) and has the potential to put low-ability and/or less-strategic test takers at a disadvantage (Aryadoust, 2012). Accordingly, low-ability listeners may fail to answer the items not necessarily due to their lack of comprehension, but due to their inability to apply the comprehension of the audio stimuli. This mechanism can be intensified by the fairly high speed of speech delivery, lengthier sentences, and higher text difficulty in Part 2 (Field, 2009). Items belonging to Part 3 are MC items with fairly lengthy stems and options. Test takers are required to read the stem and options and choose the best answer for each item, thereby engaging in reading comprehension processes.

Consequently, it appears that the test items in Parts 2 and 3 might engage additional cognitive processes besides listening comprehension and might compel learners to rely on their memory capacity and multitasking skills. To "facilitate" the listening process, the oral input is replayed for the listeners (Field, 2009). This seemingly facilitative mechanism does not improve the construct validity of the test, as it merely re-exposes listeners to the input and engages the cognitive processes; it might, however, decrease the cognitive load for some learners, as the first round of listening provides listeners with knowledge of the content and serves as a rehearsal of the test items and oral input, a process which is hardly similar to real-life listening comprehension (Field, 2009).

Class 2 listeners, who had relatively lower abilities than Class 1 listeners, were favored by fill-in-the-gap items 1, 2, 4, 5, 7, 11, and 12 in Part 1. To answer these items, test takers had to keep the stimuli in mind, summarize them into phrases no longer than three words, and supply the answer. Due to the memory demands of this task, it is expected that high-ability listeners will perform better. In addition, these items have a slower speed of delivery, fewer words and sentences, and shorter stimuli compared with Parts 2 and 3 and, therefore, are expected to favor higher-ability listeners (Buck, 2001). However, the results counter this expectation.

This unanticipated finding may be due to the items' demand for supplying answers within a word limit (three words). Aryadoust (2012) argued that high-ability learners (analogous to Class 1 in the present study) will likely consider supplying all three words as the accurate way of responding to these items, as the test instructions do not explicitly state the exact number of words to be supplied, although the word limit is set to three. Class 1 learners evidently attempted to provide lengthier responses, which is evidenced in their response sheets where many supplied three-word answers containing one or more words different from the demanded answer keys. Such semi-accurate responses are penalized in the BEC and are not awarded partial credit. This explanation is supported by the relatively large number of incorrect three-word answers supplied by Class 1 learners, but the relatively lower number of such responses by Class 2 learners.

In sum, Class 1 may be characterized as high-ability and able to comprehend oral input with medium to high text difficulty and speed of speech delivery. These learners seem to be able to multitask by keeping the stimuli in mind, reading the

test items, and choosing or supplying the answer. However, the test performance of this group might be adversely affected by the lack of clarity in the test instructions as well as the lack of partial credit in the scoring system. In contrast, Class 2 comprises low-ability listeners with limited listening and multitasking skills whose listening performance can be hampered by rapid speech containing "difficult" words.

Metacognitive Strategy Awareness, Lexico-Grammatical Knowledge, and Gender

The results showed that Class 1 had a higher average score on PS, PE, and the lexico-grammatical test, whereas Class 2 scored highly on directed attention, person knowledge, and mental translation. Using these variables, the artificial neural network classified 86% of leaners accurately into Classes 1 and 2. As expected, the most important variable characterizing latent classes was learners' lexico-grammatical knowledge, followed by problem solving, person knowledge, mental translation, planning and evaluation, and directed attention. This resonates with Imhof and Janusik's (2006) model of listening where lexico-grammatical knowledge is considered a significant factor for successful listening comprehension.

Rich vocabulary and grammar knowledge help listeners comprehend oral input (Dunkel et al., 1993); Class 1 listeners' richer and more flexible lexico-grammatical knowledge seems to have been a key element in their successful performance, especially on Parts 2 and 3. Extended lexico-grammatical knowledge facilitates automatic, less effortful listening, thereby allocating more memory space to the other tasks that listeners need to perform on the BEC (Field, 2009). Minimally proficient listeners, such as those in Class 2, with limited lexico-grammatical knowledge encounter more problems while listening to the oral stimuli; they become stuck in vocabulary recognition, likely allocating most of their memory to retrieving or surmising word meanings and applying compensatory strategies such as mental translation (Vandergrift et al., 2006).

To compensate for their deficient lexico-grammatical knowledge, less proficient listeners can spend more effort focusing their attention on the test (Goh & Hu, 2014), which may explain why their directed attention scores were higher than those of Class 1. Planning and evaluation strategies, on the other hand, are test-taking strategies that assist listeners with planning ahead and evaluating their listening performance. Class 1 had a higher mean score for planning and evaluation, suggesting that Class 1 was probably more strategy-conscious. Finally, person knowledge concerns learners' confidence and anxiety; a higher score indicates a lack of confidence in one's listening skills. As anticipated, Class 1 listeners achieved a lower person knowledge score than Class 2 listeners, suggesting that Class 1 had a lower anxiety level than Class 2. Anxiety is an affective-person factor in Imhof and Janusik's (2006) model, which is induced by lower listening ability (Vandergrift & Goh, 2012).

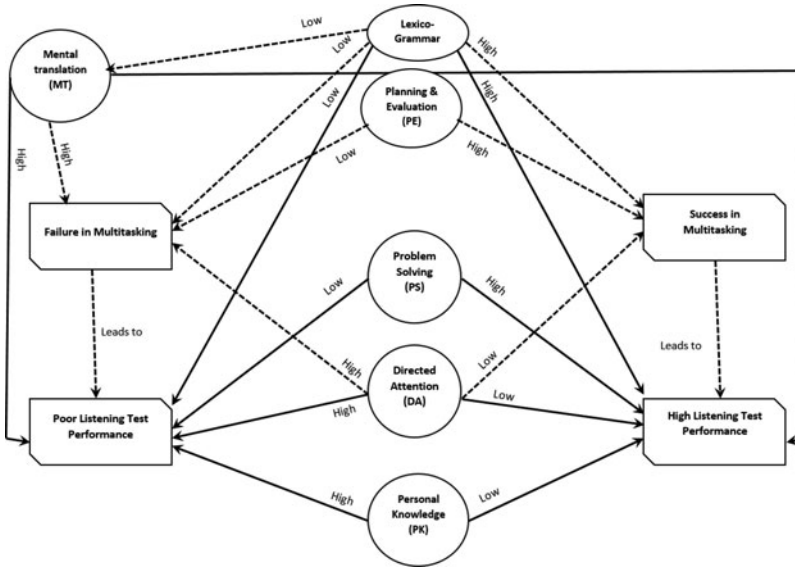

FIGURE 3
Illustration of a listening test performance model.

Unlike previous studies where gender emerged as a nuisance dimension (e.g., Aryadoust et al., 2011), the present study found no significant impact for gender. A possible explanation for this result may be the differences in the data analysis techniques. Most gender-based DIF studies used gender as a manifest variable based on which the samples were divided. The subsamples may display discrepant test patterns on a few items, but this discrepancy is not observable across all items (Chen & Jiao, 2014). In contrast, DIF emerging in latent class DIF analysis affects all or the majority of test items.

Last, whereas classification accuracy for Class 1 was significantly high across both training and testing samples (97.80% and 96%), Class 2 was classified with lower accuracy rates (25.80% and 40%), suggesting that the variance in Class 2 is likely explained by unmodeled variables. To achieve high classification accuracy for low-ability learners, using further indicators of cognitive and background factors would be useful.

A Tentative Model of Listening Test Performance

Figure 3 describes a tentative model of listening test performance based around the findings of the present study as well as previous research. The dashed lines

indicate extrapolations from previous listening research, while the solid lines represent the findings of the present study. By facilitating multitasking, high lexico-grammatical knowledge both directly and indirectly enhances listening performance; conversely, deficient lexico-grammatical knowledge directly or indirectly leads to poor listening performance by activating mental translation mechanisms. This result can be explained by the less experienced listeners' overreliance on bottom-up listening processing, which could obstruct top-down processing and the formation of a global representation of the oral input (Imhof & Janusik, 2006). Furthermore, high planning and evaluation and problem solving scores directly and/or indirectly lead to high test scores by facilitating multitasking, whereas low planning and evaluation and problem solving scores result in poor listening performance. Finally, low person knowledge and directed attention, which facilitate multitasking, appear to improve listening comprehension performance. Although lexico-grammatical knowledge, mental translation, person knowledge, and directed attention are construct-relevant factors, planning and evaluation and problem solving are largely test-specific strategies and so multitasking in this test is irrelevant to listening comprehension (Vandergrift & Goh, 2012).

Thus, we argue that the observed DIF is occasioned by both auxiliary dimensions and nuisance secondary dimensions, which largely have an adverse impact on the validity of the uses and interpretations of the test scores. This model is tentative and more research on this topic needs to be undertaken to examine the role of other elements emerging from the literature in determining test takers' success and group membership.

CONCLUSION

This article has given an account of—and possible reasons for—latent class DIF in listening comprehension, alluding to metacognitive listening strategy awareness, lexico-grammatical knowledge, and test format. The findings make several contributions to the existing literature. First, MRM analysis of DIF is argued to hold great potential for developing and validating listening comprehension models. This is the first study where MRM is applied to listening tests and a classifier artificial neural network is used for post-hoc analysis. Second, unlike DIF caused by manifest variables, latent DIF can result from numerous construct-relevant and irrelevant factors. Third, this study shows that gender may not exert a significant influence on latent DIF (although it may have a great effect when DIF is estimated based on manifest variables). Finally, this study offers some insight into the importance of metacognitive listening strategy awareness and lexico-grammatical resources, corroborating their role in determining listening skills and success in listening test performance. Further work should be undertaken to establish whether

the findings of this study are replicable across other tests administered to different groups of listeners.

Unlike traditional DIF where DIF is almost always attributed to construct-irrelevant factors, latent class DIF may be a function of both intended and unintended dimensions (Huang, Wilson, & Wang, in press). In this study, DIF was attributed to both unintended and test-related factors. For example, low-ability test takers experience test anxiety whereas high-ability test takers are not—an unintended additional dimension. Nevertheless, the high-ability group has higher lexico-grammatical scores, which is consistent with theories of listening comprehension, thus supporting the fairness of the test (Kunnan, 2013). It could be said that the latent classes result from synergy between construct-relevant and irrelevant factors. Future research is needed to take into account, for example, motivation, self-efficacy, working memory capacity, and contextual factors including formality and purpose of listening.

Finally, unidimensional extensions (e.g., the 2PL mixture item response theory model [MIRT]) and multidimensional forms (e.g., the mixture general diagnostic model) of MRM have been recently developed and applied to language assessment data with varying degrees of success (von Davier, 2008). For example, the 2PL MIRT provided useful insight in von Davier's (2005a) study of high-stakes reading and listening tests, which was comparable with multidimensional mixture models. The multidimensional discrete latent trait model (von Davier, 2005b) computer package allows for the estimation of a wide range of IRT and mixture models. More work will need to be undertaken to determine the usefulness of these models in language and educational measurement.

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.

Alexeev, N., Templin, J. L., & Cohen, A. S. (2011). Spurious latent classes in the mixture Rasch model. *Journal of Educational Measurement*, *48*(3), 313–332.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 357–374.

Aryadoust, V. (2012). Differential item functioning in while-listening performance tests: The case of IELTS listening test. *International Journal of Listening*, *26*, 40–60.

Aryadoust, V., & Goh, C. C. M. (2014). Predicting listening item difficulty with language complexity measures: A comparative data mining study. *CaMLA Working Papers*, 2014–01. Ann Arbor, MI: CaMLA.

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

Baghaei, P., & Carstensen, C. H. (2013). Fitting the mixed Rasch model to a reading comprehension test: Identifying reader types. *Practical Assessment, Research & Evaluation*, *18*(5), 1–13.

Bodie, G. D., & Crick, N. (2014). Listening, hearing, sensing: Three modes of being and the phenomenology of Charles Sanders Peirce. *Communication Theory*, *24*, 105–123.

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

Chen, Y.-F., & Jiao, H. (2014). Exploring the utility of background and cognitive variables in explaining latent differential item functioning: An example of the PISA 2009 reading assessment. *Educational Assessment*, *19*, 77–96.

Cohen, A., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, *42*, 133–148.

Dai, Y. (2013). A mixture Rasch model with a covariate: A simulation study via Bayesian Markov Chain Monte Carlo estimation. *Applied Psychological Measurement*, *37*, 375–396.

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.

Field, J. (2009). A cognitive validation of the lecture-listening component of the IELTS listening paper. In L. Taylor (Ed.), *IELTS research reports* (Vol. 9, pp. 17–65). Canberra, Australia: IELTS Australia, Pty Ltd & British Council.

Flavell, J. H., Miller, P. H., & Miller, S. A. (1993). *Cognitive development* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Frick, H., Strobl, C., & Zeileis, A. (in press). Rasch mixture models for DIF detection: A comparison of old and new score specifications. *Educational and Psychological Measurement*.

Fukuhara, H., & Kamata, A. (2011). A bi-factor multidimensional item response theory model for differential item functioning analysis on testlet-based items. *Applied Psychological Measurement*, *35*, 604–622.

Goh, C. C. M. (2000). A cognitive perspective on language learners' listening comprehension problems. *System*, *28*, 55–75.

Goh, C. C. M., & Hu, G. (2014). Exploring the relationship between metacognitive awareness and listening performance with questionnaire data. *Language Awareness*, *23*, 255–274.

Hong, S., & Min, S.-Y. (2007). Mixed Rasch modeling of the self-rating depression scale: Incorporating latent class and Rasch rating scale models. *Educational and Psychological Measurement*, *67*, 280–299.

Huang, X., Wilson, M., & Wang, L. (in press). Exploring plausible causes of differential item functioning in the PISA science assessment: Language, curriculum or culture. *Educational Psychology: An International Journal of Experimental Educational Psychology*.

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

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

Janusik, L., & Keaton, S. A. (2011). Listening metacognitions: Another key to teaching listening? *Listening Education*, *3*(2), 35–44.

Kubinger, K. D. (2005). Psychological test calibration using the Rasch model: Some critical suggestions on traditional approaches. *International Journal of Testing*, *5*, 377–394.

Kunnan, A. J. (2013). Fairness and justice in language assessment. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 540–559). Boston, MA: Wiley-Blackwell.

Li, F., Cohen, A. S., Kim, S. H., & Cho, S. J. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement*, *33*, 353–373.

Lin, P. Y., & Lin, Y. C. (2013). Examining student factors in sources of setting accommodation DIF. *Educational and Psychological Measurement*, *74*, 759–794.

Linacre, J. M. (2014). *A user's guide to WINSTEPS*. Chicago, IL: Winsteps.com.

Maij-de Meij, A. M., Kelderman, H., & van der Flier, H. (2008). Fitting a mixture item response theory model to personality questionnaire data: Characterizing latent classes and investigating possibilities for improving prediction. *Applied Psychological Measurement*, *32*, 611–631.

Oliveri, M. E., Ercikan, K., & Zumbo, B. D. (2013). Analysis of sources of latent class DIF in international assessments. *International Journal of Testing*, *13*, 272–293.

Rijmen, F., & De Boeck, P. (2005). A relation between a between-item multidimensional IRT model and the mixture Rasch model. *Psychometrika*, *70*, 481–496.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*, 271–282.

Rost, J., Carstensen, C. H., & von Davier, M. (1997). Applying the mixed Rasch model to personality questionnaires. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 324–332). New York, NY: Waxmann.

Rost, J., & von Davier, M. (1994). A conditional item fit index for Rasch models. *Applied Psychological Measurement*, *18*, 171–182.

Samuelsen, K. M. (2005). *Examining differential item functioning from a latent class perspective*. Unpublished PhD dissertation, University of Maryland, College Park.

Smit, A., Kelderman, H., & van der Flier, H. (2000). The mixed Birnbaum model: Estimation using collateral information. *Methods of Psychological Research—Online*, *5*(4), 31–43.

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

Vandergrift, L., Goh, C. C. M., Mareschal, C., & Tafaghodtari, M. H. (2006). The Metacognitive Awareness Listening Questionnaire (MALQ): Development and validation. *Language Learning*, *56*, 431–462.

von Davier, M. (2001a). *WINMIRA [Computer Software]*. Groningen, The Netherlands: ASC-Assessment Systems Corporation, USA and Science Plus Group.

von Davier, M. (2001b). *WINMIRA user manual*. Groningen, The Netherlands: ASC-Assessment Systems Corporation, USA and Science Plus Group.

von Davier, M. (2005a). *A general diagnostic model applied to language testing data* (Research Report RR-05-16). Princeton, NJ: Educational Testing Service.

von Davier, M. (2005b). *Mdltm: Software for the general diagnostic model and for estimating mixtures of multidimensional discrete latent traits models [Computer software]*. Princeton, NJ: Educational Testing Service.

von Davier, M. (2008). The mixture general diagnostic model. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 1–24). Charlotte, NC: Information Age Publishing.

Wolvin, A. D., & Coakley, C. G. (1994). Listening competency. *International Journal of Listening*, *8*, 148–160.