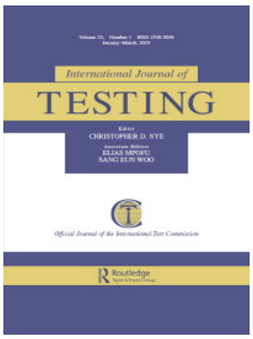

International Journal of Testing

ISSN: 1530-5058 (Print) 1532-7574 (Online) Journal homepage: www.tandfonline.com/journals/hijt20

Fitting the mixed Rasch model to the listening comprehension section of the IELTS: Identifying latent class differential item functioning

Farshad Effatpanah, Purya Baghaei, Hamdollah Ravand & Olga Kunina-Habenicht

To cite this article: Farshad Effatpanah, Purya Baghaei, Hamdollah Ravand & Olga Kunina-Habenicht (2025) Fitting the mixed Rasch model to the listening comprehension section of the IELTS: Identifying latent class differential item functioning, International Journal of Testing, 25:1, 50-89, DOI: 10.1080/15305058.2024.2414423

To link to this article: https://doi.org/10.1080/15305058.2024.2414423

Published online: 20 Oct 2024.

INTERNATIONAL JOURNAL OF TESTING
2025, VOL. 25, NO. 1, 50–89
https://doi.org/10.1080/15305058.2024.2414423

Routledge
Taylor & Francis Group

Fitting the mixed Rasch model to the listening comprehension section of the IELTS: Identifying latent class differential item functioning

Farshad Effatpanah[a], Purya Baghaei[b], Hamdollah Ravand[c] and Olga Kunina-Habenicht[a]

[a]Research Unit of Psychological Assessment, TU Dortmund University, Dortmund, Germany; [b]Department of English, Islamic Azad University, Mashhad Branch, Mashhad, Iran; [c]Department of English, Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran

ABSTRACT

This study applied the Mixed Rasch Model (MRM) to the listening comprehension section of the International English Language Testing System (IELTS) to detect latent class differential item functioning (DIF) by exploring multiple profiles of second/foreign language listeners. Item responses of 462 examinees to an IELTS listening test were subjected to MRM analysis. Three classes emerged: (1) *'Medium-level Stimulus Processors'* who can somewhat synchronize top-down and bottom-up processing, handle multitasking to a certain extent, comprehend moderately complex items, and manage input delivered at a relatively fast pace; (2) *'High-level Stimulus Processors'* who have greater abilities in synchronizing top-down and bottom-up processing, multitasking, understanding complex items, and handling fast delivery input and more paraphrased content; and (3) *'Low-level Stimulus Processors'* who rely more on bottom-up processing, have limited lexico-grammatical knowledge, struggle with multitasking and complex items, and find fast delivery input and paraphrased content challenging. Differences across the classes were further explained.

KEYWORDS

IELTS; latent class differential item functioning; L2 listening comprehension; mixed Rasch model; multiple profiles

Introduction

Listening comprehension is the ability to process, integrate, and discern implicit and explicit meaning from perceptual oral and/or visual stimuli (Buck, 2001). This skill represents a highly intricate and multidimensional cognitive process in which several (meta)cognitive and (non)linguistic skills are involved (Du & Man, 2022). Despite its pivotal role in language acquisition, production, daily communication, and academic learning (Graham, 2017), listening remains the least-researched skill, often referred to as the "Cinderella skill" in second/foreign language (L2) learning

(Field, 2013). This underrepresentation is evident in the limited research on listening test performance and the distinct cognitive processes that lead to unique listening

CONTACT Farshad Effatpanah farshad.effatpanah@tu-dortmund.de Research Unit of Psychological Assessment, TU Dortmund University, Emil-Figge Street 50, 44227 Dortmund, Germany

proficiencies (Aryadoust, 2015; Goh & Vandergrift, 2022). As a result, assessing listening comprehension ability becomes crucial across a wide spectrum of educational and professional settings, where accurate measurement can provide insight into language learners' proficiencies.

Listening comprehension tests, designed to assess examinees' understanding of spoken stimuli, are widely employed in language testing and assessment worldwide, serving as a means to measure individuals' listening proficiency. These tests are commonly integrated into a variety of standardized language proficiency assessments, each designed to fulfill distinct objectives such as placement, achievement, evaluation, and diagnostic purposes. Researchers have categorized listening tests into two types: post-listening performance (PLP) tests and while-listening performance (WLP) tests (Aryadoust, 2012, p. 41). PLP tests involve examinees listening to oral input, taking notes, and subsequently responding to a set of test items (Marx et al., 2017). Notable examples of such tests include the Michigan Language Assessment Battery (MELAB; Goh & Aryadoust, 2010) and the Test of English as a Foreign Language (TOEFL®; Isbell & Kremmel, 2020). WLP tests, by contrast, necessitate that examinees listen attentively to oral stimuli while simultaneously reading items, taking notes, selecting or composing responses, and keeping pace with the ongoing oral input stream (Aryadoust, 2012, p. 41). Prominent instances of WLP tests encompass the Certificate in Advanced English (CAE; Geranpayeh & Kunnan, 2007) and the International English Language Testing System (IELTS™; Isbell & Kremmel, 2020), developed by the University of Cambridge ESOL (English for Speakers of Other Languages) Examination Syndicate. Despite the widespread utilization of WLP tests in high-stakes standardized assessments, little attention has been devoted to such tests. The literature in this area is scant, and there is a controversy among researchers concerning the theoretical and practical aspects of these tests (Aryadoust, 2012; Field, 2009).

In most listening tests, test takers are presented with audio recordings, which can range from conversations and interviews to academic lectures or everyday spoken language scenarios. These recordings are followed by questions designed to evaluate the test takers' comprehension of the content. Test takers usually need to select the correct answer from several options, complete sentences or summaries using information from the audio, determine the truth of a statement based on what they hear, provide brief written responses based on the information from the audio, or match speakers or pieces of information to specific categories or statements. Scoring is typically binary, with responses being marked as either correct (1 point) or incorrect (0 point); partial credit is not given for incomplete or partially correct answers. The total score is calculated as the sum of all correct responses, which may then be converted into a scaled score to reflect the overall proficiency level. Furthermore, listening tests generally measure a range of cognitive skills necessary for understanding spoken language, including understanding explicit information, making inference and deduction, recognizing opinions, attitudes, or purposes, and following complex sequences. Many listening tests also report a composite score that reflects overall listening proficiency. However, some tests may report scores for specific sub-skills or dimensions of listening ability (e.g., understanding main ideas vs. specific details). These subscales, if present, are usually designed to give

more granular feedback to learners and instructors about specific areas of strength or weakness (Rost, 1990).

Scores derived from listening tests represent examinees' listening proficiency levels and furnish valuable evidence for making informed decisions about each examinee. When interpreting the results of listening tests, test users and developers must prioritize the validity of these tests—ensuring that they are appropriate for their intended purposes and uses (He & Jiang, 2020). This emphasis on validity is of paramount importance as the inferences and decisions drawn from test scores bear consequences for all stakeholders involved. Consequently, concerns pertaining to the quality and reliability of listening tests remain prevalent within the field of language assessment research.

In the field of educational testing, two main statistical models are used to scale examinees. These models are the Item Response Theory (IRT; Embretson & Reise, 2000) and the Rasch model (Rasch, 1980). These models determine the position of examinees' scores along a single continuous proficiency scale. Subsequently, these scores are utilized to facilitate comparisons or rank-ordering of examinees against specific criteria or in relation to their peers. The primary assumption underlying psychometric models is that individual differences among examinees are *quantitative* variations (Rost, 1990). While it is true that all examinees employ varying degrees of the same strategies and skills to answer a set of test items, it is important to recognize that they may employ these strategies and skills in distinct ways or even opt for entirely different solution patterns or strategies. This indicates the presence of *qualitative or structural* variations among examinees as well (Rost, 1990). Consequently, test scores reflect not only quantitative differences in a single construct but also represent the qualitative differences in the types of strategies, skills, or processes that examinees adopt to give a correct response to a given test item. Without acknowledging these qualitative differences, examinees would essentially be compared across different test constructs using the same test (Baghaei et al., 2019).

A statistical approach that proves invaluable in specifying the test construct and, notably, in identifying both quantitative and qualitative distinctions among examinees is the analysis of Differential Item Functioning (DIF; Holland & Wainer, 1993). DIF methods are used to ascertain whether different subgroups with the same levels of ability exhibit varying responses to specific test items. When subgroups employ different strategies and processes in addressing test items, observed DIF signals the presence of qualitative differences among them. As noted by De Ayala and Santiago (2017), these differences may stem from variations in language or educational backgrounds, self-concept, and diverse test-taking experiences, among other factors. The occurrence of DIF calls into question the assumption of unidimensionality and raises doubts about the credibility of test score interpretations and uses. Most significantly, DIF signifies that the test is measuring distinct constructs across subgroups, rendering the ranking or comparison of examinees along the same proficiency continuum inappropriate.

Given this background, the primary objective of this study is to employ the Mixed Rasch Model (MRM; Rost, 1990) in the analysis of the listening comprehension section of the IELTS, a well-established standardized WLP test. The aim is to investigate latent class DIF by exploring multiple profiles of L2 listeners who are likely to employ diverse listening comprehension processes in order to correctly respond to a set of test items.

Background

Listening comprehension and multiple profiles

Numerous researchers have proposed various models aimed at elucidating the intricacies of the listening comprehension process and its relationship with a plethora of (non)cognitive attributes. As delineated by Aryadoust (2018), these models can be broadly categorized into two groups. The first group encompasses general models that primarily focus on listening within non-assessment contexts. For instance, models such as those proposed by Imhof and Janusik (2006), Rost (2016), and Goh and Vandergrift (2022) posit that listening involves a multifaceted cognitive journey that commences with the pre-comprehension phase, characterized by processes of perception and recognition. During perception, auditory organs receive sound waves, which are subsequently converted into electrical impulses and transmitted to the brain for processing *via* the nervous system. Recognition, the second process, involves the identification and retrieval of phonemes and words through lexical access and segmentation. In the comprehension stage, syntactic knowledge is applied in a bottom-up fashion (i.e., the use of sounds, individual words, and smaller units to construct meaning from the auditory stimuli) to amalgamate the identified words. This amalgamation results in the creation of a localized mental representation of perceived chunks or sentences (Kintsch, 1998). These representations contain occasional gaps and are comprised of propositions and mental imagery. The task of storing these mental representations falls upon long-term memory, as listeners decode, analyze, and encode the received chunks or sentences. When a new stimulus is encountered, a similar comprehension process unfolds, connecting the new input to the previously stored mental representation of the listening stimuli. This connection is facilitated through a top-down process (i.e., the use of linguistic, contextual, pragmatic, and sociolinguistic knowledge to facilitate auditory comprehension) that draws upon one's background knowledge and the formulation of inferences. This process enables listeners to bridge gaps and construct a coherent mental representation referred to as the situation model (Aryadoust, 2018). Both bottom-up and top-down processes rely on the listener's memory as their synchronization constrains working memory capacity and impacts test performance (Buck, 2001).

Researchers have indicated that cognitive processing occurs in an interactive rather than a linear manner and depends not only on linguistic knowledge but also on world and topical knowledge (Field, 2013; Goh & Vandergrift, 2022). More importantly, various levels of processing are required to comprehend an oral input. Lower-level cognitive processes (e.g., lexico-grammatical knowledge, word recognition, parsing, and acoustic-phonetic decoding) are engaged to grasp the literal meaning of a text, whereas higher-level processes (e.g., semantic processing, world knowledge resources, inferencing, multitasking, and speakers' intentions and prosodic patterns) are activated to comprehend the discourse and implied meaning of an input (Field, 2013; Rukthong & Brunfaut, 2020). Previous studies have shown that automatic or fluent use of lower-level processes reduces the cognitive processing load, enabling listeners to allocate more attentional capacity to higher-level processes (Goh & Vandergrift, 2022).

The second group encompasses models that are specifically tailored for assessment purposes (Bejar et al., 2000; Buck, 2001; Buck & Tatsuoka, 1998; Field, 2013; Freedle

& Kostine, 1996). These models take into account both the default listening mechanisms and a range of test- and test-taker-related characteristics (Aryadoust, 2018). Among these assessment-specific models, one of the most influential models has been presented by Bejar et al. (2000), which delineates listening comprehension into two stages: the listening stage and the response stage. During the listening stage, verbal stimuli are received by the listener's auditory system and subsequently processed. To comprehend incoming signals, listeners must have real-time access to at least three knowledge sources: (1) Situational knowledge, which underscores the significance of contextual knowledge and visual cues in aiding listening comprehension, (2) Linguistic knowledge, encompassing grammar (phonology, vocabulary, morphology, and syntax), discourse, and pragmatics, and (3) Background knowledge, pertaining to one's knowledge of the world and the current situation. Throughout this stage, the oral input is transformed into a set of propositions. Owing to variations in knowledge, cognitive processing, and working memory capacity among different listeners, various sets of propositions may be generated (Bejar et al., 2000).

In the response stage, the incoming input is processed to formulate a response, which can take the form of spoken or written expression (e.g., selecting an option, filling in a blank space, or providing words or phrases). Bejar et al. (2000) also emphasize the role of several factors in listening test performance and the interconnectedness of these factors. These factors include task characteristics (e.g., delivery speed, task complexity, exposure to listening input, discourse register, genre, audio file repetition, test materials, and rubric), individual characteristics (e.g., age, gender, and background knowledge), and linguistic and cognitive knowledge (e.g., lexico-grammatical knowledge, concentration, and working memory capacity).

Furthermore, alongside these models, numerous researchers have posited that listening comprehension comprises various subskills and have proposed various taxonomies (e.g., Field, 2013; Richards, 1983; Rost, 2016, as comprehensively reviewed by Aryadoust, 2018). While these models and taxonomies have yielded valuable insights into the nature of listening comprehension and provided implications for pedagogy, assessment, and research, they often assume that listeners exhibit similar processes and patterns when processing listening input and responding to test items. This assumption implies homogeneity among listeners in terms of listening processes. However, it is important to recognize that individual listeners may differ in their listening processes, and there may be multiple viable configurations through which listeners achieve successful listening comprehension. As posited by Wolvin (2013, p. 105),

> the individual differences of listeners might take us to a reconceptualization of listening in the broader communication context. Rather than a 'one size fits all' model of the listening process, perhaps we should focus on individual listeners' processing strategies. It may be that listeners vary considerably in their cognitive functioning while engaging as listening communicators.

Moreover, Hickendorff et al. (2018) assert that total raw scores are typically employed in statistical analyses to develop models. The use of total scores assumes homogeneity in response patterns, with any heterogeneity within and between individuals considered primarily as statistical noise.

Therefore, only a limited number of researchers have adopted a profiling analysis approach and designed scales to measure individual differences by identifying styles, habits, and patterns of listening, and categorizing listeners into different profiles. For instance, Watson et al. (1995) developed a sixteen-item Listening Styles Profile (LSP-16) inventory to assess intra-individual variability in listening styles. They defined listening styles as "attitudes, beliefs, and predispositions about the how, where, when, who, and what of the information reception and encoding process" (p. 2). Using factor analysis, they derived four distinct patterns of listening styles: content-oriented, time-oriented, action-oriented, and people-oriented listening styles. Addressing a common criticism of the LSP-16 regarding its consistently low estimates of internal consistency (Bodie & Worthington, 2010), Bodie et al. (2013) proposed a revised 24-item measure of LSP (LSP-R). Employing exploratory factor analysis (EFA), they investigated the underlying factor structure of the measure, which assesses four listening styles: task-oriented, analytical, relational, and critical listening. In another study, Bodie et al. (2020) first developed a typology of listening habits based on facets of meaning construction and subsequently designed a corresponding measure to capture individual differences. Utilizing EFA, they identified four distinct listening habits: Connective, Reflective, Analytical, and Conceptual. Chon and Shin (2019) also conducted a study to theorize and substantiate intra-individual differences in students' motivational-metacognitive profiles regarding their listening proficiency. Using latent class analysis (LCA), they identified four cluster solutions: Amotivated-Translators, Externally Motivated-Don't do much Planning or Evaluation, Introjected-Totally Alert, and High Autonomous Motivation-Achievement Strategists. This was based on the responses of 312 Korean middle school learners of English to a metacognitive awareness listening questionnaire and an academic self-regulation questionnaire.

While these studies have shown the presence of different profiles of listening styles and processes, they are subject to several limitations. First, these studies have often employed inadequate methodologies for exploring listening profiles. Early studies relied on factor analysis methods, while recent ones have tended to use latent class approaches. Factor analysis methods are limited in their ability to accurately describe heterogeneity and complex, non-linear listening patterns. According to Hickendorff et al. (2018, p. 2), factor analysis methods and other traditional analytical approaches, including correlation, regression-based techniques, and Analysis of Variance (ANOVA), are variable-centered and focus on relationships between variables. They assume that the relationship between variables applies uniformly to all individuals, suggesting homogeneity in the nature of individual differences (Hickendorff et al., 2018). Consequently, they are ill-suited for providing a clear representation of non-linear and interactive patterns and addressing heterogeneity within and between individuals. Latent class approaches, such as LCA, also rely on observed variables and mean scores (Tabachnick & Fidell, 2013) and are incapable of modeling item responses (Aryadoust, 2015). Second, these empirical studies have primarily focused on intra-individual differences to characterize listening patterns. Third, these studies were conducted under non-assessment conditions. In assessment contexts, if qualitatively different groups exist within a population, it serves as empirical evidence for the presence of DIF.

All in all, the research discussed above underscores the complexity and variability of listening comprehension processes, highlighting that different listeners exhibit distinct cognitive patterns and strategies. While previous studies have explored these individual differences through various profiling approaches, there remains a gap in effectively capturing the heterogeneity of listening processes. In the context of language assessments like IELTS, such variability can manifest as DIF, where certain test items may function differently for different groups of listeners. Consequently, further research is needed to detect DIF and better account for the diverse listening profiles that exist within the population that enhance our understanding of how individual differences impact assessment outcomes.

Differential item functioning (DIF)

DIF is identified when individuals with the same level of the construct being measured, but from different predefined groups (such as age, gender, ethnicity/race, education, etc.), have different probabilities of endorsing an item (Zumbo, 2007). Essentially, DIF can be seen as a form of measurement bias where examinees' responses to test items are influenced by factors beyond the primary construct the test intends to measure (Ravand, 2015; Roussos & Stout, 1996). In other words, item responses are not solely determined by the primary construct but also by group membership, indicating the presence of multidimensionality. This situation can threaten the validity of test interpretations and uses (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014).

Within the context of IRT and the Rasch model, DIF detection methods involve comparing areas between item response functions (Raju, 1990) or item difficulty parameters (Thissen et al., 1993). To maintain unidimensionality, test scores or item difficulty[1] estimates should be invariant across groups of the population with similar ability levels. When difficulty estimates for items differ among groups, it implies that examinees from different groups employ distinct cognitive processes and strategies to respond to items, revealing that group membership influences test performance.

Several statistical methods are available for detecting DIF, including Mantel-Haenszel (Holland & Thayer, 1988), logistic regression (Swaminathan & Rogers, 2000), multiple-group factor analysis (Meredith, 1993; Ravand, 2024), multiple indicator multiple cause (MIMIC; Finch, 2005; Ravand et al., 2019), IRT-/Rasch-based analytical methods (Raju, 1988; Steinberg & Thissen, 2006), and multidimensional IRT (MIRT; Oshima et al., 1997). These methods typically examine DIF by analyzing the manifest characteristics of examinees (e.g., age, gender, ethnic group, race, etc.). However, this approach may not always identify the root causes of DIF effectively, as it focuses on researcher-defined characteristics and may overlook unidentified characteristics, leading to heterogeneity among emerging groups (Cohen & Bolt, 2005; Geranpayeh & Kunnan, 2007).

[1]DIF is not necessarily limited to item difficulty because it could also be applied to item discrimination (e.g., Humphry & Montuoro, 2021; Lord, 1980). Although the focus on item difficulty fits with the use of the Rasch model, it must be pointed out that other characteristics of the item can vary as well.

As argued by Ackerman et al. (2003), psychometric DIF analysis should always be followed by a substantive investigation of the sources of DIF, bridging the gap between statistical and substantive analyses. While statistical analysis is crucial for validity, understanding the underlying causes of DIF is enlightening from both theoretical and substantive perspectives, especially in terms of construct validation (Borsboom et al., 2004; Van Nijlen & Janssen, 2011).

DIF in L2 listening comprehension

DIF analysis in listening comprehension research has conventionally been conducted using several covariates such as gender, age, grade, nationality, place of origin, academic background, first language or language background, prior exposure to similar tests (test-wiseness or practice effect), familiarity with topic, and familiarity with item type (e.g., Aryadoust, 2012; Geranpayeh & Kunnan, 2007; Seo et al., 2016). Among the covariates, gender has been the most well-researched factor in the DIF literature on listening assessment. In previous L2 listening comprehension studies conducting DIF analyses, gender was either the sole variable examined (e.g., Aryadoust et al., 2011; Lin & Wu, 2003; Park, 2008) or one of several variables investigated for its potential impact on differential test performance (e.g., Aryadoust, 2012; Cid et al., 2017; Geranpayeh & Kunnan, 2007; Seo et al., 2016). However, the findings of the studies for the presence of DIF were inconclusive and contradictory. A group of studies reported that males outperformed females, or males were disadvantaged on some items and advantaged on others (Alavi et al., 2018; Aryadoust, 2012; Aryadoust et al., 2011; Cole, 1997; Park, 2008; Zansen et al., 2022). By contrast, another group of studies found that females outperformed their male counterparts (Lin & Wu, 2003). Some studies also reported no gender-based DIF of practical concern (Bourdeaud'Hui et al., 2021; Cid et al., 2017). These mixed findings are sensible because the studies used different tests and detection methods for DIF analyses. Additionally, previous studies indicated that nationality and age do not induce DIF, but grade, place of origin, familiarity with topic and item type, prior exposure to tests, academic background, and first language can cause DIF (Aryadoust, 2011, 2012; Banerjee & Papageorgiou, 2016; Geranpayeh & Kunnan, 2007; Harding, 2012; Lia & Yao, 2021; Nishizawa, 2023; Pae, 2004; Raquel, 2019; Shin et al., 2021). Some researchers further focused on investigating the interaction of several covariates in inducing DIF (e.g., Aryadoust et al., 2024; Pae, 2012).

Researchers have explored various hypotheses to explain differential performance on listening items in standardized tests, with a focus on gender differences. Park (2008) attributed observed DIF in the English listening subtest of the Korea College Scholastic Ability Test to factors such as item content, topic areas, and language type, noting that items about travel and sports favored males, while those about theater and shopping favored females. Other studies echoed these findings, showing items favoring females typically relate to arts and social sciences, while those favoring males involve natural sciences and technical content (Carlton & Harris, 1992; Curley & Schmitt, 1993; O'Neill & McPeek, 1993; Scheuneman & Gerritz, 1990). Females also excel in computations and symbolic items, whereas males perform better on geometry and items involving tables and graphs (O'Neill & McPeek, 1993). Additionally, females tend to outperform

males on items involving mood, contextual clues, or abstract concept understanding, while males do better on logical inference tasks (O'Neill & McPeek, 1993). Studies also indicate that females excel in oral and constructed response formats (e.g., essay-type, short answers, and fill-in-the blanks) due to stronger writing skills, while males often perform better on multiple-choice (MC) items, partly due to a greater willingness to guess (Aryadoust, 2012; Bolger & Kellaghan, 1990; Mazzeo et al., 1993; Pae, 2012; Willingham & Cole, 1997). Cognitive processing differences also contribute, with females showing stronger verbal skills and males relying more on spatial processing, affecting performance across various task types (O'Neill & McPeek, 1993). For instance, males excel in spatial tasks like map labeling, which involve visualization, real-life information, and spatial understanding, while females perform better in tasks involving linguistic analysis, detailed comprehension, and linguistic inference such as sentence and table completion (Aryadoust, 2012). Table 1 provides descriptions of abilities required for different listening comprehension item types, adapted from Buck (2001) and Aryadoust (2012).

Although previous DIF studies explain the causes of DIF and characterize the differential performance of examinees on listening comprehension tests, only a few studies have specifically focused on WLP tests, particularly the IELTS exam (e.g., Alavi et al., 2018; Aryadoust, 2012). Focusing on WLP tests is crucial because it ensures that the

Table 1. Listening comprehension item types and their description.

Item type	Description
Multiple-choice	- Ability to attentively listen to comprehend the main idea and details of oral stimuli. - Ability to identify main information, such as main points, supporting details, and inferred meanings. - Ability to understand context and use inference skills to deduce implied meanings. - Ability to distinguish between similar options and eliminate distractors.
Map labeling	- Spatial awareness and visualization skills to understand and interpret maps. - Ability to follow the sequence of directions and comprehend spatial relationships between locations. - Ability to accurately identify and label locations on a map based on verbal descriptions. - Ability to attentively listen for specific details related to geographical features, landmarks, and directions.
Sentence completion	- Ability to understand the context and meaning of sentences. - Ability to predict and infer missing information based on preceding content. - Ability to recognize keywords and phrases that appropriately complete missing information. - Knowledge of grammar, vocabulary, sentence structure, phonology, and morphology to accurately complete sentences.
Table completion	- Ability to extract and comprehend information presented in tables. - Ability to identify patterns and relationships within the table. - Ability to attentively listen for specific details related to numerical data, categories, or attributes (e.g., numbers, dates, names, etc.). - Ability to retain and accurately transfer information from the oral input to the table.
Matching items	- Ability to understand relationships between items presented in the oral stimuli. - Ability to recognize similarities and differences between items. - Ability to categorize and classify information based on shared characteristics. - Ability to identify key information and making connections between items. - Ability to retain and manipulate information in memory while processing the task. - Ability to follow the sequence of information presented and retain the order in memory. - Ability to infer the correct matches based on given information.
Classification	- Ability to categorize and group information based on shared characteristics. - Ability to identify similarities and differences between items. - Ability to understand the overall structure and organization of information presented. - Ability to recognize patterns and relationships to accurately classify items. - Ability to retain, remember, and retrieve specific details from the oral stimuli.

assessments accurately reflect the real-time processing abilities of examinees, which is an essential component of effective listening comprehension. The limited attention to DIF analysis in WLP ESOL tests may be attributed to Cambridge University's standards that prioritize score consistency over other aspects of test quality, such as fairness and test validity (Geranpayeh & Kunnan, 2007).

Mixed Rasch model

The MRM (Rost, 1990) is a combination of the Rasch model (Rasch, 1980) and the latent class model (LCM; Lazarsfel & Henry, 1968). The Rasch model assumes that the probability of getting an item right depends on the ability of a person and the difficulty of an item. The greater a person' ability relative to an item difficulty, the higher is the probability of a right answer. The position of an item on a latent variable continuum corresponds with the position of person at which there is a 0.5 probability of a correct response to the item. The probability of a correct response is dependent on the difference between the ability of a person and the difficulty of an item. Generally, when the difference is negative, the probability of responding correctly to the item is less than 0.50; when the difference is 0, the probability is equal to 0.50; and when the difference is positive, the probability is greater than 0.50. For the standard Rasch model, the item response function is expressed as:

$$P(X_{vj}=1|\theta_v.\beta_j)=\frac{exp\left(\theta_v-\beta_j\right)}{1+exp(\theta_v-\beta_j)} \tag{1}$$

where $P(X_{vj})$ denotes the probability of solving item j for examinee v; θ_v is the ability of person v; and β_j is the difficulty of the item j.

The Rasch model relies heavily on the assumption of parameter invariance, which asserts that the difficulties of items should remain constant across all members of the population. In fact, the ordering of item difficulties should be equal for all individuals. However, this assumption might not hold in cases where there are qualitative distinctions, such as variations in cognitive approaches, among different (sub)groups. To mitigate this issue, the MRM relaxes the assumption by permitting item parameters to vary across latent population classes.

The assumption of MRM is that the population is heterogeneous but involves various non-overlapping subpopulations that are different in terms of item response probabilities. The MRM can be expressed as:

$$P(X_{vj}=1|\theta_{vg}.\beta_{jg}.g_v)=\frac{exp\left(\theta_{vg}-\beta_{jg}\right)}{1+exp\left(\theta_{vg}-\beta_{jg}\right)} \tag{2}$$

where $P(X_{vj})$ is the probability of a correct response; θ_{vg} is the ability parameter of person v in class g; β_{jg} is the class specific difficulty parameter of item j in class g; and g_v is a latent class person v belongs to it. It must be noted that unlike the LCM, which

assumes that there are no individual differences within each class regarding the response probabilities, the MRM allows for individual differences within latent classes. Therefore, while the unidimensional Rasch model holds within each latent class, it does not hold for the entire population. Since latent classes are a priori unknown and disjoint, each person must belong to only one latent class with the highest probability (Rost, 1990).

There are typically two approaches to the analysis of MRM: (1) exploratory analysis approach and (2) confirmatory analysis approach. In the exploratory approach, a number of latent classes are firstly detected, and then class-specific item profiles are analyzed to understand the nature of differences among the latent classes. This entails a thorough examination of the content and patterns of variation in item parameters within each class. Such scrutiny can illuminate qualitative differences among the classes for researchers. The association between several covariates and the latent class membership can also be explored to elucidate the qualitative distinctions among these classes. In contrast, the confirmatory approach incorporates several covariates directly into the model while estimating the latent classes (De Ayala & Santiago, 2017). The presence of substantial a priori evidence aids in confirming whether the covariates moderate person and item parameters (Sen & Cohen, 2019).

The MRM has already been used in methodological and practical studies to test fit of the unidimensional Rasch model, standard setting, test calibration, DIF, test speededness, problem-solving strategies, and response styles and faking personality (see Sen & Cohen, 2019 for a comprehensive review of applications of IRT mixture models). The application of MRM to listening comprehension tests has received too little attention. Most directly, Aryadoust (2015) applied the MRM to an EFL listening comprehension test. In addition to the listening test, respondents were given a metacognitive awareness listening questionnaire and a lexico-grammatical test. Two latent groups were detected. The first class consisted of examinees with higher ability in multitasking and lexico-grammatical knowledge, and with higher scores in planning, evaluation, and problem solving; however, the second class comprised examinees with lower ability in multitasking and lexico-grammatical knowledge, and with lower scores in mental translation, person knowledge, and directed attention. The class analysis revealed that examinees in Class 1 outperformed examinees in Class 2 on matching items (items for which examinees must select the accurate options from a list of choices and write the letters next to the item numbers). Aryadoust (2015) argued that these concurrent cognitive-motor tasks are extraneous to listening comprehension. Although the study provided valuable information about individual differences in listening test performance, there was a priori hypothesis for using the MRM. A study with a priori hypotheses concerning the causes of DIF using MRM is not a reasonable strategy. The purpose of MRM is to identify DIF across latent classes that are otherwise unknown, and no a priori assumption is required to detect DIF across the classes (Baghaei et al., 2019). Results of MRM would not be informative when researchers first perform MRM and then attempts to make an association between the classes and covariates that have been chosen a priori based on a theory. This method can be utilized when analysis of class-specific item profiles fails to produce interpretable results. When researchers have covariates that are speculated to be the causes of DIF, conventional DIF detection techniques for known groups can be used to analyze DIF for these variables (Baghaei et al., 2019; Rost, 1990).

The present study

The present study aims to apply the MRM (Rost, 1990) to the listening comprehension section of the IELTS to investigate latent class DIF by exploring multiple profiles of L2 listeners, focusing on examining the patterns of item difficulty parameters across latent classes and analyzing the content of test items to capture qualitative differences among them. For the purpose of this study, the following research questions were addressed:

RQ1: How many latent classes (qualitative profiles in listening mechanism) do exist among examinees in taking IELTS listening comprehension items?

RQ2: Which items exhibit a significant difference in item difficulty estimates across latent classes?

RQ3: What are the distinct cognitive abilities required for effectively answering various types of listening comprehension items, and how do these abilities differ across various question types?

Method

Data

The present study capitalized on a dataset that had been previously utilized by Aryadoust (2012) and Effatpanah (2019). The data consisted of item responses of 462 international students to forty items of the listening comprehension section of a sample paper of the IELTS. Participants were studying in language schools and tertiary-level institutions in Iran, Malaysia, Singapore, and the Philippines. There were 191 (41.34%) males and 271 (58.66%) females, and their mean age was 24.46 ($SD = 4.27$). Participants were from Iran (76.62%), China (11.91%), Malaysia (5.84%), and from other countries, mostly Arab states in the Persian Gulf region (5.63%). They were preparing for the IELTS exam and their participation was voluntary. A written informed consent had been obtained from all individual participants included in the study. They had been reassured that their information would remain confidential and anonymous. They also had received complete feedback on their performance. The feedback contained their raw scores, IELTS band scores, information regarding their deficiencies in listening comprehension, and suggestions for improving their listening skills.

Due to confidentiality reasons, the version of the IELTS listening test had been taken from Official IELTS Practice Materials (2007) (www.IELTS.org,). Just similar to other IELTS versions, this test was subjected to a rigorous test design, development, and validation process which rely on the seven-stage Cambridge ESOL Question Paper Production Cycle (CEQPPC). This process not only assures that every version of the test is of a comparable level of difficulty but also provides evidence for the plausibility of the interpretations and uses of the test scores. The listening test usually lasts for a total of 40 minutes. The recordings are heard only once and include a range of accents, including British, American, Australian, New Zealand, and Canadian. The items are designed so that the answers appear in the order they are heard in the audio. One mark is awarded for each correct answer, and care should be taken when writing answers on the answer sheet because poor spelling and grammar are penalized. There are no partially correct answers; each answer is marked as either 1 (correct) or 0 (incorrect). For

the paper-based listening test, after the recording ends, examinees are given 10 minutes to transfer their answers to the answer sheet.

The test that had been administered to the participants comprised four sections (four audio stimuli), each with ten questions. A variety of question types were used, including multiple-choice (MC), map labeling (ML), sentence completion (SC), and table completion (TC). No sample items are provided in this article due to copyright constraints. Readers can refer to the IELTS website (n.d.). The first section of the test consisted of two ML items, four MC items, and four TC items on a woman being interviewed by a police officer. The second section was composed of five MC items and five TC items on providing commercial information about an English Hotel. The third section included one MC item and nine SC items based on a conversation among three students on campus. The last section comprised six SC items and four TC items based on a talk presented by a university lecturer on a bird of prey. The total score in the test ranged between 4 and 40 with a mean of 21.32 and standard deviation of 9.12. Reliability coefficients of the test were estimated using Cronbach alpha, and a value of 0.92 was obtained.

Data analysis

The item responses of all examinees were analyzed with the MRM using the *TAM* package (Robitzsch et al., 2024) in R Core Team (2024). The *TAM* package uses marginal maximum likelihood estimation (MMLE) and joint maximum likelihood estimation (JMLE) methods for unidimensional and multidimensional IRT models as well as dichotomous and polytomous models. Most of conventional (parametric) DIF detection methods are based on the premise that a certain group of items, called anchor items, are free from DIF. These anchor items serve to create a constant scale for comparing item difficulty across different (sub)groups (Glas & Verhelst, 1995). The presence of suitable anchor items is critical for accurate DIF analysis. The violation of this assumption can result in inflated type I error rate and erroneous inferences (Wang et al., 2022; Yuan et al., 2021). The *TAM* package uses the equal-mean-difficulty anchor method for DIF analysis. This method assumes that the mean of difficulty across items is the same across classes or (sub)groups (typically equal to zero). Therefore, in this study, the mean of difficulty across items was constrained to be equal to zero for each latent class in the MRM analysis.

Several researchers (e.g., Alexeev et al., 2011; Sen, 2018) have indicated that when a test aligns with, for instance, the two or three parameters logistic (2PL or 3PL) IRT models, utilizing an MRM might lead to the detection of spurious latent classes, which, in turn, can lead to erroneous or ambiguous conclusions that have considerable effects on practitioners. Therefore, it is essential to first examine whether the tests align with the 2PL and 3PL IRT models (Tay et al., 2011), unless there exists a valid rationale why these models would not yield as much valuable insight as MRM. The estimation of the 3PL IRT model with an additional guessing parameter appeared not reasonable for the data used in this study due to the multi-item format of the listening test, including MC and fill-in-the-gap items. Thus, a 2PL mixture IRT model was fitted to the data and its results were compared against the MRM.

Since the number of classes is not a model parameter to be estimated, Rasch models with one to five latent classes were fitted to the data and compared to explore the optimal number of classes. Their results were also compared against the 2PL mixture IRT models with one to five latent classes. The models were compared with regard to relative fit statistics: Akaike's Information Criterion (AIC) = $-2\ log\ L + 2P$, where L is the maximum likelihood function value, P is the number of parameters; Bayesian Information Criterion (BIC) = $-2\ log\ L + P\ \ln[n]$, where ln[n] is the natural log of sample size; and Bozdogan's Consistent AIC (CAIC) = $-2\ log\ L + p\ [\ln(n) + 1]$. The model with the least information criteria is considered the best model. Burnham and Anderson (2002) argued that AIC asymptotically selects the model that reduces the mean square error prediction. With a simulation study, Li et al. (2009) also indicated that AIC is less accurate and inconsistent unless the true model is among the rival models. However, because AIC does not impose any penalty for sample sizes, the more complex or highly parameterized model is selected with an increase in sample size. Numerous researchers, on the contrary, have shown that BIC has the superiority to detect the correct number of latent classes because it imposes a large penalty for the number of parameters and sample sizes (Choi et al., 2017; Li et al., 2009; Nylund et al., 2007; Preinerstorfer & Formann, 2012; Sen et al., 2019), so BIC tends to choose models with a smaller number of parameters compared to AIC.

The model identification was followed by estimating item difficulty and person parameters as well as examining mean square (MNSQ) fit statistics for each latent class. Item difficulty and person ability parameters with logit units or log odds units for test items/persons indicate the location of each item/person on the latent trait continuum (e.g., L2 listening comprehension ability). Class-specific item difficulty parameters were also graphically compared to identify the items that cause qualitative differences between classes. The mean and standard deviation of the weighted likelihood estimate (WLE) of person parameters for latent classes were also computed. WLE person parameter estimates show the ability scores on the Rasch scale for latent classes after controlling for the item difficulty pattern difference. To detect further differences between the resulting groups, they were compared with regard to their reliability, means of raw scores across latent classes, and the correlation of item difficulties across the classes, along with their confidence intervals (CIs).

Furthermore, the fit of individual items for each latent class was tested using outfit and infit MNSQ fit indices (Linacre, 2002), their z-standardized values, and p-value. According to Linacre (2002), infit MNSQ is a t-standardized information-weighted statistic which is sensitive to inliers, whereas outfit MNSQ is a t-standardized unweighted statistic which is sensitive to outliers. ZStd provides a t-test to examine whether the data have a perfect fit to the model. A p-value is used to test the statistical significance of the observed misfit. The acceptable range for infit and outfit MNSQs is $0.70 - 1.30$ (Bond et al., 2020; Linacre, 2024). Overall, larger values of infit MNSQ show that the items of a test do not perform well for the examinees on whom the items are targeted, suggesting a more serious threat to validity (Linacre, 2024).

Finally, to further identify which items had a significant difference in item difficulty across latent classes, a post hoc t-test was conducted. The Welch t-test was used to assess whether the differences between the difficulty estimates based on each class are statistically significant. In this method, item difficulty parameters are separately

estimated for each group through logistic regression. The difference between these estimates is then tested for statistical significance. DIF items are identified when $p < 0.05$ (Linacre, 2024). When DIF is both statistically and substantively significant and consistently replicates across various subgroups, the researcher can more confidently conclude that the item functions differently across these subgroups (Aryadoust, 2012). The formula for the Welch *t*-test is as follows:

$$t = \frac{d_{j2} - d_{j1}}{\sqrt{s_{j2}^2 - s_{j1}^2}} \tag{3}$$

where d_{i1} denotes the difficulty of item j for group g, and s_{i1}^2 is the standard error of estimate for item j for group g. Researchers at Educational Testing Service (ETS) developed a standardized metric, known as the ETS DIF classification scheme (Holland & Thayer, 1988), to categorize DIF into three levels based on effect size (Δ): negligible DIF (A) if $|\Delta| \leq 1$, intermediate DIF (B) if $1 \leq |\Delta| \leq 1.5$, and large DIF (C) if $|\Delta| \geq 1.5$. These categories help identify whether test items function differently across different groups after controlling for the overall ability level.

Results

To explore the appropriate number of latent classes, several Rasch models with one to five latent classes were fitted to the data, and their results were compared against those of 2PL mixture IRT models with different latent classes. Table 2 provides the relative model fit statistics across the models. As can be seen, information criteria (e.g., AIC and BIC) did not produce consistent results. Such discrepancies between information criteria have been frequently reported in previous studies for model comparisons (Sen & Cohen, 2019). Nylund et al. (2007) argued that compared to AIC, BIC has a higher probability of selecting the correct number of latent classes in factor mixture modeling. With regard to AIC, the value of the 2PL mixture IRT model with five latent classes was the smallest due to its higher complexity. However, the values of BIC showed that 2PL mixture IRT models had a poor fit compared to the Rasch models with different latent classes, except for 2PL mixture IRT model with one latent class that outperformed the one-latent class Rasch model. This suggests that the 2PL item response functions do

Table 2. Model-data relative fit information for the estimated MRMs and 2PL mixture IRT models.

Models	AIC	BIC
Rasch One-Latent Class	18752	18922
Rasch Two-Latent Class	18309	18652
Rasch Three-Latent Class	18071	18587
Rasch Four-Latent Class	17907	18598
Rasch Five-Latent Class	17779	18643
2PL One-Latent Class	18294	18625
2PL Two-Latent Class	17954	18620
2PL Three-Latent Class	17741	18741
2PL Four-Latent Class	17519	18855
2PL Five-Latent Class	17459	19130

Table 3. Class-specific statistics across the three-latent class model.

Class	Class size	Mean probability class 1	Mean probability class 2	Mean probability class 3
1	0.509	0.961	0.037	0.002
2	0.278	0.036	0.943	0.021
3	0.213	0.006	0.042	0.952

not accurately capture the underlying structure of the data. The BIC value for a three-latent-class Rasch model was the smallest, and thus this model was determined to be the most effective for the sample in this study and chosen for subsequent analyses. This indicates that assuming three homogeneous subgroups offers a more accurate representation of the data compared to assuming a single population.

The class size decimals for each latent class are presented in Table 3. The sum of these sizes is equal to one, and they are interpreted as percentages, suggesting the number of examinees assigned to be members of each latent class. About 50%, 28%, and 21% of the examinees were identified as members of Classes 1, 2, and 3, respectively. Table 3 also gives the mean hypothetical class assignment probabilities of the three latent classes. For instance, members of Class 1 had a very high probability of being classified in Class 1 (96.1%) and a low probability of assignment to Class 2 (3.7%) and Class 3 (0.2%). Members of Class 2 had a chance of 94.3% to be in Class 2; the chances to be in one of the other two classes were 3.6% and 2.1%, respectively. Similarly, members of Class 3 had a probability of 95.2% of being in Class 3, and they had 4.2% and 0.6% probabilities of being classified in Classes 2 and 1, respectively. Three well-separated classes were thus clearly identified. The off-diagonal indices on the probability matrix are much smaller than the diagonal statistics, indicating a high classification accuracy of the model (Baghaei & Carstensen, 2013; Effatpanah et al., 2024).

The mean and standard deviation of WLE of person parameters for the three latent classes were computed. The mean and standard deviation of person parameters for Class 1 were 1.379 and 0.941; for Class 2 they were 1.485 and 0.832; and for Class 3 they were −1.055 and 0.800. This suggests the higher ability of Class 2 members. Class 1 had the highest standard deviation, indicating greater variability in listening comprehension abilities among its members.

Table 4 shows the difficulty parameters for the items, their standard errors, and their infit and outfit MNSQ values. As can be seen, although most infit and outfit values were within the acceptable range of 0.70–1.30 (Bond et al., 2020; Linacre, 2024), there were some misfitting items across the three classes. For instance, the outfit values for Items 29 (0.69), 36 (1.95), and 38 (0.67) in Class 1 were out of the acceptable boundary, but their misfit was insignificant. For Class 2, Items 3 (1.38), 9 (1.61), 14 (1.75), 15 (1.41), 22 (1.34), and 37 (0.65) were beyond the criteria. Misfit was significant only for Item 3. And, for Class 3, although twelve items (i.e., 2, 3, 5, 8, 9, 10, 13, 14, 17, 21, 24, and 34) fell out of the expected range of outfit MNSQ, only Item 5 had a significant misfit. Items with values lower than 0.70 are overfit and benign. However, items with values exceeding 1.30 indicate abnormal response patterns that deviate from the model's expectations, and suggest that the test is not unidimensional. A possible reason for this can be due to the presence of several item types in the listening section of the test that can potentially affect the unidimensionality of the test. When different item types are introduced in a listening test, they may require different cognitive processes. For example, MC items might assess recognition and selection skills, while constructed response

Table 4. Item difficulty parameters and item fit statistics for the three latent classes.

	Class 1				Class 2				Class 3			
Items	Estimate	S.E.	Infit MNSQ	Outfit MNSQ	Estimate	S.E.	Infit MNSQ	Outfit MNSQ	Estimate	S.E.	Infit MNSQ	Outfit MNSQ
1	−1.50	0.37	0.92	0.86	−2.98	0.69	1.00	0.81	−1.66	0.15	1.01	0.76
2	−2.55	0.59	0.95	0.92	−3.65	0.87	1.01	0.96	−1.70	0.15	0.93	0.22
3	0.88	0.20	1.10*	1.13*	−2.24	0.48	1.23*	1.38*	−0.85	0.14	0.99	1.63
4	−0.41	0.25	1.04	1.09	0.57	0.22	0.98	0.98	−0.21	0.15	1.03	0.98
5	0.29	0.22	1.03	1.03	2.16	0.24	1.02	1.06	−0.70	0.14	1.24*	1.36*
6	2.11	0.20	1.02	1.06	−0.49	0.27	0.94	0.95	0.91	0.18	1.18	1.09
7	−1.81	0.42	0.90*	0.88*	−1.81	0.43	0.97	1.00	−0.60	0.14	1.06	0.87
8	−1.81	0.42	1.03	1.01	−2.56	0.59	0.97	1.06	−2.98	0.20	1.07	1.68
9	−1.25	0.34	1.06	1.06	−2.98	0.71	1.12	1.61	−2.34	0.16	0.89	0.25
10	0.24	0.22	0.99	1.00	−0.91	0.31	1.01	0.99	−0.29	0.14	1.10	1.41
11	0.80	0.20	1.05	1.07	0.20	0.23	0.99	0.92	−0.34	0.14	0.97	0.94
12	−2.96	0.67	1.09	1.12	−0.42	0.27	1.02	1.16	−1.66	0.15	0.93	0.85
13	−2.96	0.71	1.03	1.18	−3.68	0.93	1.01	0.96	−2.06	0.16	0.98	0.37
14	−1.14	0.32	0.93	0.92	−1.81	0.43	1.09	1.75	−1.99	0.15	1.07	1.86
15	−0.41	0.25	0.94	0.93	−0.21	0.25	1.07	1.41	−1.17	0.14	1.01	0.91
16	−0.22	0.24	1.08	1.12	−1.81	0.43	1.05	1.18	−1.29	0.14	1.07	1.29
17	0.76	0.20	0.98	0.95	3.03	0.28	0.99	0.96	0.07	0.15	0.79	0.58
18	−0.77	0.28	0.98	0.96	−0.03	0.24	1.03	1.12	−0.27	0.14	1.00	1.22
19	−0.28	0.25	1.13*	1.18*	−0.35	0.26	1.07	1.24	−1.59	0.14	0.98	0.89
20	−0.22	0.24	0.93	0.88	0.15	0.24	1.04	1.04	0.52	0.16	1.07	1.09
21	1.11	0.20	1.01	0.94	3.31	0.31	1.04	1.04	1.50	0.22	1.18	1.39
22	−0.11	0.23	0.90	0.85	0.72	0.22	1.11	1.34	0.14	0.15	1.01	1.01
23	−0.17	0.24	0.97	0.83	1.68	0.22	0.97	0.89	1.45	0.22	0.97	0.95
24	2.28	0.21	0.99	1.02	3.13	0.30	0.96	1.06	3.26	0.45	0.76	0.53
25	−0.41	0.25	1.05	1.05	−0.08	0.25	0.99	0.97	0.05	0.15	0.97	0.82
26	1.57	0.19	1.04	1.02	3.52	0.33	0.99	0.98	1.45	0.21	1.04	1.15
27	1.30	0.19	0.92	0.93	−1.57	0.34	0.93	0.94	1.50	0.22	0.95	0.71
28	−1.81	0.41	1.03	1.01	−2.24	0.48	0.98	0.91	−0.02	0.15	1.00	1.07
29	−0.76	0.27	0.97	0.69	1.74	0.23	1.05	1.08	2.76	0.34	1.20*	1.18

30	−0.85	0.29	1.02	0.93	−0.21	0.25	1.01	0.93	0.42	0.16	1.06	1.01
31	1.53	0.19	1.01	1.00	1.25	0.22	0.99	0.99	−0.91	0.14	1.03	1.02
32	0.60	0.20	0.96	0.84	0.15	0.24	0.92	0.89	1.24	0.20	0.87	0.80
33	−0.13	0.23	0.97	0.94	−0.57	0.28	0.89	0.77	0.01	0.15	1.02	0.89
34	1.34	0.19	1.01	1.12	−0.49	0.27	0.91	0.89	1.08	0.19	0.81	0.62
35	0.47	0.21	1.03	1.03	−0.82	0.30	0.86	0.79	0.75	0.17	0.99	0.82
36	0.76	.020	1.04	1.95	2.00	0.23	0.86	0.81	3.25	0.42	0.80*	0.73*
37	−0.94	0.30	0.99	0.99	−1.35	0.36	0.92	0.65	−0.42	0.14	1.06	0.94
38	2.45	0.21	1.00	0.67	1.68	0.22	0.96	0.96	4.69	0.70	0.98	0.94
39	1.95	0.20	0.95	0.74	0.09	0.24	0.92	0.89	2.53	0.32	1.02	0.97
40	1.34	0.19	0.95	0.79	0.86	0.22	1.09	1.08	1.55	0.22	1.02	0.98

Note.
*$p < 0.05$; S.E. = Standard Error of measurement; MNSQ = Mean-Square.

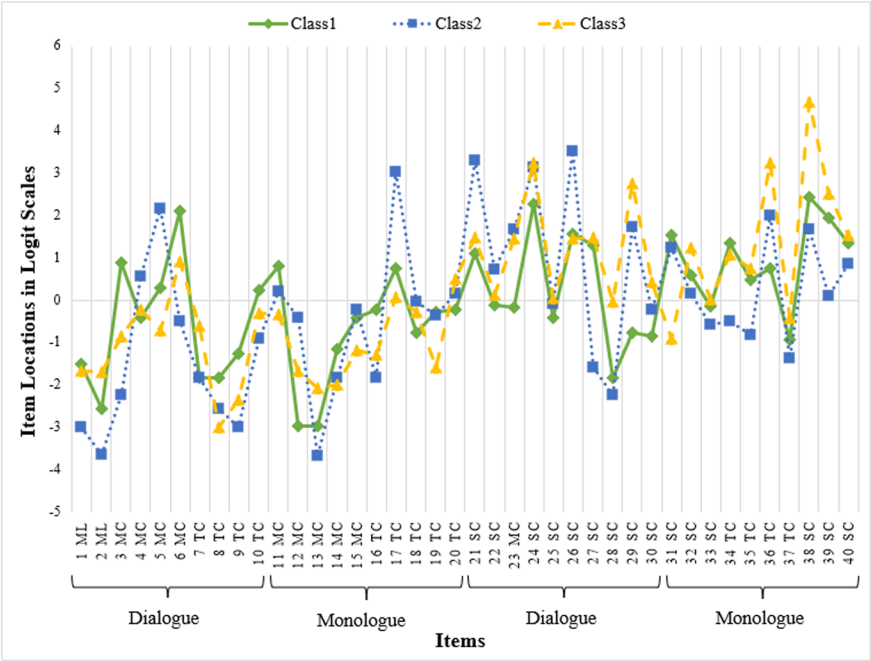

Figure 1. Item difficulty parameters across the three-latent classes.

items (e.g., short answers) might assess production and elaboration skills. These differences can introduce multiple dimensions to the test.

The pattern of class-specific item difficulty parameters across the three latent classes is graphically presented in Figure 1. The vertical axis shows difficulty estimates of the items in logit scale, and the horizontal axis indicates 40 items of the test. Class 1 is depicted with a solid green line, Class 2 with a dotted blue line, and Class 3 with a dashed yellow line. The patterns of difficulty parameters showed an inconsistent pattern across the three classes, suggesting different cognitive structures for examinees from the three classes (Effatpanah et al., 2024; Rost, 1990). That is, there are substantial qualitative differences between examinees with regard to their listening comprehension process. Item difficulty parameter estimates for Class 3 showed greater variability (ranging from −2.98 to 4.69) than Classes 1 and 2 in which item difficulty parameter estimates ranged from −2.96 to 2.45 and −3.68 to −3.52, respectively. The easiest items for Class 1 were Items 12, 13, and 2, while Items 13, 2, 1, and 9 were the easiest items for Class 2, and Items 8, 9, and 13 for Class 3. On the other hand, Items 38, 24, and 6 were the most difficult items for Class 1, whereas the most difficult items for Class 2 were 26, 21, and 24, and Items 38, 24, and 36 for Class 3. As illustrated in Figure 1, almost all items contributed to variability among the three latent classes, especially Items 3, 4, 5, 6, 12, 17, 29, 36, and 38.

To detect which items exhibit significant differences in difficulty parameters across the classes and which items have consistent difficulty levels across them, post hoc *t*-tests on the difficulty parameter differences for each item across the classes were conducted. Table 5 presents the local difficulty contrasts across classes, the standard error of the

Table 5. Welch *t*-test in the difficulty parameter difference for each item across the three classes.

	Class 1 vs. Class 2					Class 1 vs. Class 3					Class 2 vs. Class 3						
Items	DIF Contrast	Joint S.E.	*t*	*p*-value	ETS	DIF Contrast	Joint S.E.	*t*	*p*-value	ETS	DIF Contrast	Joint S.E.	*t*	*p*-value	ETS	Item Format	Type of Speech
1	1.48	0.78	1.89	0.060	B	0.16	0.40	0.40	0.689	A	−1.32	0.70	−1.87	0.063	B	ML	Dialogue
2	1.11	1.05	1.05	0.294	B	−0.85	0.60	−1.40	0.163	A	−1.95	0.88	−2.21	0.029*	C	ML	
3	3.12	0.52	6.00	0.000*	C	1.74	0.24	7.17	0.000*	C	−1.39	0.50	−2.76	0.006*	B	MC	
4	−0.98	0.34	−2.89	0.004*	A	−0.20	0.29	−0.68	0.500*	A	0.78	0.27	2.93	0.004*	A	MC	
5	−1.87	0.32	−5.86	0.000*	C	0.99	0.26	3.85	0.000*	A	2.86	0.27	0.41	0.000*	C	MC	
6	2.60	0.34	7.73	0.000*	C	1.20	0.27	4.43	0.000*	B	−1.40	0.33	−4.30	0.000*	B	MC	
7	0.00	0.60	0.00	0.998	A	−1.21	0.44	−2.73	0.007*	B	−1.21	0.45	−2.70	0.008*	B	TC	
8	0.75	0.73	1.02	0.308	A	1.17	0.47	2.50	0.013*	B	0.42	0.62	0.67	0.501	A	TC	
9	1.74	0.79	2.20	0.029*	C	1.10	0.38	2.92	0.004*	B	−0.64	0.73	−0.87	0.386	A	TC	
10	1.15	0.38	3.06	0.003*	B	0.54	0.26	2.07	0.040*	A	−0.61	0.34	−1.80	0.073	A	TC	
11	0.60	0.31	1.94	0.053	A	1.14	0.25	4.63	0.000*	B	0.54	0.28	1.96	0.051	A	MC	Monologue
12	−2.54	0.72	−3.52	0.001*	C	−1.30	0.69	−1.89	0.060*	B	1.24	0.30	4.07	0.000*	B	MC	
13	0.71	1.17	0.61	0.542	A	−0.90	0.73	−1.24	0.216	A	−1.61	0.94	−1.72	0.088	C	MC	
14	0.67	0.54	1.26	0.210	A	0.86	0.36	2.41	0.017*	A	0.18	0.46	0.40	0.691	A	MC	
15	−0.20	0.36	−0.55	0.586	A	0.76	0.29	2.61	0.010*	A	0.95	0.29	3.28	0.001*	A	MC	
16	1.58	0.49	3.22	0.002*	C	1.06	0.28	3.79	0.000*	B	−0.52	0.45	−1.16	0.248	A	TC	
17	−2.27	0.34	−6.66	0.000*	C	0.69	0.25	2.74	0.007*	A	2.96	0.31	9.41	0.000*	C	TC	
18	−0.74	0.38	−1.98	0.049*	A	−0.49	0.32	−1.55	0.122	A	0.25	0.28	0.87	0.384	A	TC	
19	0.06	0.36	0.17	0.862	A	1.31	0.28	4.59	0.000*	B	1.25	0.30	4.16	0.000*	B	TC	
20	−0.37	0.34	−1.10	0.272	A	−0.75	0.29	−2.55	0.011*	A	−0.37	0.29	−1.29	0.197	A	TC	
21	−2.20	0.36	−6.04	0.000*	C	−0.38	0.29	−1.31	0.191	A	1.81	0.38	4.81	0.000*	C	SC	Dialogue
22	−0.83	0.32	−2.57	0.011*	A	−0.26	0.28	−0.91	0.363	A	0.57	0.27	2.14	0.034*	A	SC	
23	−1.85	0.33	−5.67	0.000*	C	−1.62	0.32	−5.05	0.000*	C	0.23	0.31	0.75	0.454	A	MC	
24	−0.85	0.36	−2.36	0.020*	A	−0.99	0.50	−1.98	0.049*	A	−0.13	0.54	−0.25	0.804	A	SC	
25	−0.33	0.36	−0.94	0.351	A	−0.46	0.30	−1.56	0.120	A	−0.13	0.29	−0.44	0.658	A	SC	
26	−1.96	0.38	−5.13	0.000*	C	0.12	0.29	0.40	0.688	A	2.07	0.39	5.29	0.000*	C	SC	
27	2.88	0.39	7.35	0.000*	C	−0.19	0.29	−0.67	0.506	A	−3.07	0.40	−7.60	0.000*	C	SC	
28	0.43	0.63	0.68	0.498	A	−1.79	0.44	−4.08	0.000*	C	−2.22	0.50	−4.41	0.000*	C	SC	
29	−2.49	0.35	−7.11	0.000*	C	−3.52	0.43	−8.12	0.000*	C	−1.02	0.41	−2.51	0.012*	B	SC	
30	−0.64	0.39	−1.65	0.100	A	−1.27	0.33	−3.80	0.000*	B	−0.63	0.30	−2.08	0.038*	A	SC	

(*Continued*)

Table 5. Continued.

	Class 1 vs. Class 2					Class 1 vs. Class 3					Class 2 vs. Class 3						
Items	DIF Contrast	Joint S.E.	*t*	*p*-value	ETS	DIF Contrast	Joint S.E.	*t*	*p*-value	ETS	DIF Contrast	Joint S.E.	*t*	*p*-value	ETS	Item Format	Type of Speech
31	0.28	0.29	0.97	0.333	A	2.44	0.24	0.21	0.000*	C	2.16	0.26	8.33	0.000*	C	SC	Monologue
32	0.45	0.31	1.44	0.152	A	−0.64	0.29	−2.23	0.027*	A	−1.09	0.31	−3.51	0.001*	B	SC	
33	0.44	0.36	1.21	8.229	A	−8.13	8.28	−8.48	0.633	C	−0.57	0.32	−1.81	0.072	A	SC	
34	1.83	0.33	5.49	0.000*	C	0.26	0.27	0.95	0.343	A	−1.57	0.33	−4.74	0.000*	C	TC	
35	1.29	0.36	3.53	0.001*	B	−0.28	0.27	−1.03	0.303	A	−1.57	0.35	−4.53	0.000*	C	TC	
36	−1.23	0.31	−4.01	0.000*	B	−2.49	0.47	−5.35	0.000*	C	−1.25	0.48	−2.61	0.009*	B	TC	
37	0.41	0.47	0.87	0.383	A	−0.52	0.33	−1.56	0.119	A	−0.93	0.39	−2.41	0.017*	A	TC	
38	0.77	0.31	2.49	0.014*	A	−2.24	0.73	−3.06	0.002*	C	−3.01	0.74	−4.09	0.000*	C	SC	
39	1.86	0.31	5.99	0.000*	C	−0.58	0.38	−1.53	0.127	A	−2.44	0.40	−6.08	0.000*	C	SC	
40	0.48	0.29	1.63	0.105	A	−0.21	0.30	−0.70	0.487	A	−0.68	0.31	−2.18	0.030*	A	SC	

Note. S.E.: standard error of measurement.
*$p < 0.05$; ETS: educational testing service; A: negligible; B : moderate; and C: large.

DIF contrasts, the Welch t value, the p-value for the contrasts, and ETS DIF classification of effect size measures. The 'DIF Contrast' columns provide the difference between the local difficulty estimates of the items across the classes. A positive DIF contrast indicates that the item is more difficult for the first, left-hand-listed class, and a negative DIF contrast indicates that the item is more difficult for the second, right-hand-listed class. The Joint S.E. is the standard error of the DIF contrast. The Welch t value also shows the statistical significance between the local difficulties of items as a Student's two-sided t statistic (Linacre, 2024). The null hypothesis is that the two estimates are the same, except for measurement error. For instance, as illustrated in Table 4, the difficulty of Item 1 is −1.50 for Class 1 and −2.98 for Class 2; the contrast in difficulty is 1.48 with a joint SE of 0.78, as depicted in Table 5, indicating that Item 1 is more difficult for members of Class 1; the Welch t value of this contrast is 1.89; and the p-value of the contrast is 0.060, which does not meet the established significance threshold of 0.05. The DIF contrast value shows a moderate (B) DIF.

As can be seen in Table 5, the DIF analysis between Classes 1 and 2 identified 22 items (i.e., 3, 4, 5, 6, 9, 10, 12, 16, 17, 18, 21, 22, 23, 24, 26, 27, 29, 34, 35, 36, 38, and 39) with significant DIF at $p<0.05$. The analysis between Classes 1 and 3 also revealed that most items function differentially across the classes, and fifteen items (i.e., 1, 2, 13, 18, 21, 22, 25, 26, 27, 33, 34, 35, 37, 39 and 40) did not exhibit differential functioning between the two classes. Similarly, the analysis between Classes 2 and 3 indicated that only 14 out of 40 items did not function differently across the classes (i.e., 1, 8, 9, 10, 11, 13, 14, 16, 18, 20, 23, 24, 25, and 33). Items 3, 4, 5, 6, 12, 17, 29, 36, and 38 showed differential functioning across the three classes.

The substantial difference across the classes was supported by a moderate Spearman rank-order correlation between difficulty parameter estimates of Classes 1 and 2 ($r=0.605$, 95% CI [0.353–0.775], $p<0.001$), of Classes 1 and 3 ($r=0.697$, 95% CI [0.485–0.831], $p<0.001$), and of Classes 2 and 3 ($r=0.676$, 95% CI [0.454–0.819], $p<0.001$). This suggests slight agreement between the item parameter estimates across the classes. The mean difference in raw scores can be used to represent the difference in listening comprehension ability across classes because most items for the three classes fitted the Rasch model well, and the mean of item difficulty was assumed to be consistent across the classes to link the metrics of classes to a common scale (Rasch, 1977; Wang, 2004). Independent-sample t-tests were thus performed to investigate whether mean of raw scores across the three classes were statistically significant. The results showed a significant difference in mean between Class 2 ($M=29.63$, $SD=5.000$) and Class 1 ($M=27.77$, $SD=5.026$, t (222) = 2.748), $p=0.006$), suggesting that members of Class 2 showed considerably better test performance compared to Class 1 members. There was also a significant difference in mean between Class 2 and Class 3 ($M=13.84$, $SD=5.436$, t (363) = 27.177), $p<0.001$) as well as between Class 1 and Class 3 (t (333) = 21.744), $p<0.001$). This indicates that members of Class 1 had a significantly better performance relative to Class 3 members and that Class 2 members significantly outperformed members of Class 3. A minor difference in reliability was also observed across the three latent classes. For Class 1, the reliability was 0.79, with 99% CIs ranging from 0.75 to 0.83. Class 2 showed a reliability of 0.79, with CIs between 0.71 and 0.82, while Class 3 had a reliability of 0.79, with CIs of 0.73–0.85.

Discussion

This study applied the MRM to examine latent class DIF in the listening comprehension section of the IELTS. The goal was to identify multiple profiles of listeners who exhibit qualitative differences in their listening processes when answering test items. Alternative Rasch and 2-PL models with one to five latent classes were considered, and the Rasch model with three latent classes yielded the best fit. Class 1 comprised approximately 50% of the sample, while Classes 2 and 3 included about 28% and 21%, respectively. To capture qualitative differences among the classes, some speculations regarding the individual differences of the classes are first proposed. The processes examinees of the three classes may distinctively use to answer the test items are characterized. Then, a content analysis is conducted to identify potential causes of DIF in test items.

Labeling and characterizing latent classes

The emergence of three latent classes indicates significant qualitative differences in how examinees in each class approached the listening test items. The results of person parameters and mean test performance across the classes revealed that Class 2 consists of high-level examinees with the highest listening ability and less variability; Class 1 includes moderate-level examinees with high proficiency but more variability; and Class 3 is comprised of low-level examinees with the lowest ability and moderate variability. Most of the items generally favored examinees in Class 2, although the items from Section 3 mostly functioned in favor of examinees in Class 1. The easiest items for the three groups belonged to Sections 1 and 2, while the most difficult items for Classes 1 and 3 were from Sections 3 and 4. Examinees in Class 2 found items in Section 3 as the most challenging items.

Overall, the results appear to emphasize that examinees with varying levels of listening proficiency approach the items differently. In fact, there is a different pattern in the information processing capacity of examinees with different listening abilities. To comprehend listening input and produce correct answers, L2 examinees need to utilize both lower- and higher-level processing skills. As the L2 listening literature suggests (Field, 2013; Rukthong & Brunfaut, 2020), it is unlikely that examinees can successfully activate higher-level cognitive processes without effectively engaging in lower-level cognitive processing, such as lexico-grammatical knowledge, word recognition, and parsing. Examinees must utilize higher-level cognitive processes to understand the main point of the input they are listening to (Rost, 2016). Therefore, higher-level examinees use both types of cognitive processes to answer test items. As articulated by Goh and Vandergrift (2022), examinees with higher-level listening abilities can seamlessly synchronize the higher- and lower-level as well as top-down and bottom-up processes in a rapid, almost subconscious manner. However, low and moderate-ability examinees are dependent on "controlled listening processes which entail conscious attention to and processing of elements in the speech stream." (p. 19). Fluent or automatic accessing of lower-level skills takes up little of examinees' attention, thereby leaving sufficient cognitive capacity for higher-level skills (Rost, 2016). This aligns with Cognitive Load Theory (Sweller, 1988), which emphasizes the importance of managing cognitive load to maximize learning efficiency. More particularly, fluent access to linguistic repertoire significantly enhances the

efficient use of working memory (Field, 2013). This efficient use of working memory is crucial, as Limited Attentional Capacity theory suggests that finite attentional resources must be allocated effectively to process information successfully (Kahneman, 1973).

Therefore, it seems that Class 2 includes examinees who can effectively activate both top-down and bottom-up processes as well as lower- and higher-level skills. They are also efficient at decoding sounds, words, and syntactic structures due to their high listening proficiency and greater lexico-grammatical knowledge. Moderate-level examinees (Class 1) seems to possess both some ability in either top-down or bottom-up processes, but lack proficiency in integrating both effectively, and a decent grasp of lower-level skills, but struggle with higher-level comprehension tasks. Also, they are generally good at decoding sounds and words, but may occasionally struggle with more complex or less familiar linguistic structures, and can parse syntax and grammar effectively, though perhaps not as quickly or accurately as the high-level group. However, due to their limited linguistic repertoire, low-level examinees (Class 3) appears to struggle with recognizing sounds, words, and syntactic structures, leading to difficulties in constructing appropriate meaning from oral input, misinterpreting or missing parts of the stimuli, and hindering overall comprehension. More importantly, members of this class are likely to depend on bottom-up listening processing which may potentially hinder their ability to effectively engage in top-down processing and develop a comprehensive mental representation of the auditory stimuli (Imhof & Janusik, 2006). In essence, examinees in Class 3 exhibit a lower aptitude for coordinating between top-down and bottom-up processes and retrieving pertinent knowledge from memory compared to their counterparts in Classes 2 and 1. Any difficulty in retrieving lower-level processes also increases the cognitive processing load and prevent a listener from leaving sufficient cognitive capacity for higher level processes. This finding also aligns with Limited Attentional Capacity theory (Kahneman, 1973) stating that individuals have limited attentional resources, and the involvement of higher-level subskills likely increases the cognitive load. Class 3 examinees may have struggled to allocate their attention effectively between the low- and high-level skills, indicating their lack of competence in distributing their attentional resources across the skills more effectively.

Given their limited lexico-grammatical knowledge, examinees in Class 3 seem to struggle with vocabulary recognition and use most of their memory capacity to recover word meanings (Aryadoust, 2012). Consequently, these examinees are more likely to employ specific strategies to mitigate the adverse impact of working memory limitations. Several researchers have posited that individuals with lower to moderate listening abilities employ metacognitive strategies and compensatory mechanisms (e.g., relying on general world knowledge, common sense, mental translation, cultural information, and visual, contextual, or paralinguistic cues) to manage their listening processes. These strategies help compensate for their deficiency in specific subskills of the target language and facilitate the coordination between lower-level and higher-level processing (Effatpanah, 2019; Goh & Vandergrift, 2022). Echoing this perspective, Harding et al. (2015, p. 12) contend that "comprehension does not strictly adhere to a linear progression from lower-level to higher-level processing; instead, various levels may operate concurrently, with difficulties at one level being offset by 'positive information' at another, or with simultaneous challenges at both higher and lower levels leading to overall miscomprehension."

The study also highlighted the impact of genre on test performance, with Sections 1 and 2 focusing on general topics and Sections 3 and 4 dealing with academic contexts. Genre has been found to significantly affect examinees' performance (Chen & Chen, 2021), emphasizing the importance of considering genre-related factors in listening comprehension tests. Previous studies showed that although genre is not an important factor for grasping the theme of the text, it has a great impact on understanding the details and main points of lectures (Chen & Chen, 2021). The results revealed that examinees across the classes had better performance in the two first sections relating to general topics, while examinees in Class 2 with higher listening comprehension ability outperformed in the last two sections focusing on academic places.

In the later sections of the IELTS listening test, items become more complex, including more paraphrased content, longer sentences, multitasking tasks, and faster delivery. Generally, items in Sections 1 and 2 mainly involve understanding smaller chunks of oral input and require lower-level cognitive processing, such as grasping factual information and functional relationships. In contrast, items in Sections 3 and 4 involved understanding longer chunks of oral input and demanded higher-level cognitive processing, such as making inferences, paraphrasing, and integrating listening skills with other abilities like reading, note-taking, and writing (Aryadoust, 2012; Effatpanah, 2019). It appears that the better performance of examinees in Class 2 can emanate from their ability to handle more challenging items and possess more cognitive capacity and stimulus-focused attention for increasing their speed of processing. However, certain factors like lengthy sentences and unclear item instructions may inadvertently tap into reading comprehension processes and negatively affect the performance of examinees. Class 3 examinees struggled in Section 4, which required them to follow a fast-paced stream of oral input and integrate listening ability with other skills. Integrated tasks like lectures in Section 4 are more authentic but demanding. Examinees in Class 3 may have lagged behind the audio stream, leading to missed items. It has been shown that integrated tasks such as the lecture type in Section 4 of the IELTS test are more authentic than conventional item formats, including MC and matching items, and reduce the effect of background knowledge (Rukthong & Brunfaut, 2020). However, Aryadoust (2012) argues that the concurrent exposure to written and oral inputs impedes note-taking. Therefore, it can be assumed that those examinees who fell behind the stream of oral stimuli missed several questions. This can be attributed to restricted reading skills, memory capacity, stimulus-focused attention, test wiseness, test-taking strategies, and other confining factors (Aryadoust, 2012; Estaji & Banitalebi, 2023; He & Jiang, 2020).

Taken together, the three distinct classes of listeners can be labeled as: "*High-level Stimulus Processors*" (Class 2); "*Moderate-level Stimulus Processors*" (Class 1); and "*Low-level Stimulus Processors*" (Class 3). Class 2 examinees exhibited better abilities in synchronizing top-down and bottom-up processing, operationalizing higher-level cognitive processes, understanding longer chunks of oral input, possessing more cognitive capacity and stimulus-focused attention, handling multitasking and integrated items, comprehending complex items, and understanding items with a high speed of delivery and paraphrased content. Class 1 examinees exhibited a balanced approach to listening tasks. They are capable of effectively combining top-down and bottom-up processing but may not do so as seamlessly as Class 2. These examinees can understand and interpret oral input well but may require a bit more effort to comprehend longer or more complex

stimuli. They demonstrate good cognitive capacity and can manage multitasking and integrate items reasonably well. However, their performance may vary depending on delivery speed and paraphrased content. Class 3 examinees, in contrast, struggled with these tasks, relying more on lower-level processing and metacognitive strategies to compensate for their limitations. These findings shed light on diverse approaches individuals take in listening comprehension tests and highlight the impact of various cognitive processes on test performance.

Content analysis of test items

The post-hoc analysis and the item difficulty profiles across the three classes indicated that almost all items contributed to the differences among the classes. A content analysis of the items was conducted to identify the causes of observed DIF and further analyze the above-mentioned speculations about the information processing of examinees across the classes. The identification of main sources of DIF is often demanding, especially in exploratory DIF studies where a priori hypothesis is absent (Zumbo, 2007). The authors consulted previous studies (i.e., Aryadoust, 2012; Effatpanah, 2019; Geranpayeh & Taylor, 2008) in which attributes required to correctly answer listening items of the IELTS were discussed. Test items primarily tap examinees' linguistic knowledge, understanding of detailed information, comprehending explicitly stated general and literal information, understanding of paraphrases, making inferences, and comprehending of illocutionary meaning. Geranpayeh and Taylor (2008) argue that the listening inputs, developed by the University of Cambridge ESOL Examination Syndicate for WLP tests, are designed "with some internal repetition", and that test items focus on "explicit and easily accessible information" to decrease "any potential negative impact of hearing the text only once in slightly adverse conditions", with "key information rephrased and repeated within the text" to allow examinees to confirm their answers as they listen (p. 3). Therefore, such tests appear to narrowly reflect the listening construct by concentrating mainly on the understanding of details (Aryadoust, 2012), which impose severe demands on the examinees' memory load as they need to focus on specific details and memory skills (Shohamy & Inbar, 1991). In the following sections, the potential causes of DIF across the four sections of the test are discussed.

Section 1

Items 1 and 2. In this section (social dimension), there is a dialogue between a woman and a police officer in which the woman tells the story of a robbery. The attributes or primary dimension targets for these items are linguistic knowledge, world knowledge, making inferences, ability to understand detailed or specific factual information, and ability to integrate listening with visual skills. Our post-hoc analysis showed that Item 2 was significantly easier for Class 2 members than Class 3. They were generally expected to outperform on this item because they can more accurately identify locations and effectively follow instructions (i.e., spatial and directional understanding) due to their higher listening comprehension ability. The significant difference across the classes could also be due to the confounding role of gender. Previous studies reported that males tend to perform better on map labeling items than females because of their higher verbal processing capacity (Aryadoust, 2012; O'Neill & McPeek, 1993).

Items 3–6. These items significantly contributed to variability among examinees across the three classes. The primary dimensions are linguistic knowledge, world knowledge, ability to make paraphrase, the ability to understand detailed or specific factual information, and making inferences. As expected, Items 3 and 6 favored Class 2 with the lowest item difficulties. However, Items 4 and 5 functioned in favor of low- and medium-level examinees (Classes 3 and 1). This unexpected result accords with previous studies reporting that examinees whose listening comprehension ability ranges from low to moderate level tend to achieve higher scores on MC items (Aryadoust, 2012; Chang & Read, 2013). It could be attributed to the influence of test-taking strategies, with low- and moderate-ability listeners being more inclined to guess, and these guesses sometimes result in correct answers. With three options for each MC question (Items 3 to 5), there is a relatively high probability of success (33%) when making a random guess.

Furthermore, research on cognitive psychology has shown that males exhibit a propensity to take greater risks, such as opting for a lucky guess, when confronted with a problem (Buck, 2001). Consequently, males tend to adopt a more risk-taking approach, potentially resulting in higher scores and excel over females in MC items (Bolger & Kellaghan, 1990; Mazzeo et al., 1993). However, females exhibit greater reluctance to guess on MC items compared to males and are inclined to skip items they are uncertain about (Aryadoust, 2012). Another possible reason is that high-level examinees might miss easy items due to carelessness as there is a lack of shared incorrect answers among Class 2 members. This conjecture finds reinforcement in the outfit MNSQ patterns of this Class: numerous incorrect responses by high-ability examinees on easy items had outfit MNSQ values exceeding 1.3, indicating that their performance on these items was unanticipated.

Items 7–10. These items measure examinees' linguistic knowledge, world knowledge, ability to understand detailed or specific factual information, and ability to integrate listening, reading, short-term memory span, and writing abilities. Our post-hoc analysis indicated that except for Item 8, the other items functioned in favor of Class 2, with the smallest item difficulty values. However, the results showed some unexpected patterns. For example, the advantage of Item 8 for Class 3 was unexpected. Items 9 and 10 also indicated a disadvantage for Class 1 compared to Class 3. Given that outfit MNSQ is sensitive to outliers, this suggests that some high- and moderate-level examinees missed easy items, and low-level examinees correctly answered more difficult items. Moreover, the differences across the classes could be attributed to the effect of gender. Research has indicated that females have better performance on table completion items because these items require linguistic inference and detailed comprehension (Aryadoust, 2012).

Section 2

Items 11–15. In this section (social dimension), there is a woman providing commercial information about an English Hotel (Bridge Hotel). The primary dimensions are linguistic knowledge, world knowledge, ability to make paraphrases, ability to understand detailed or specific factual information, and make inferences. The results of post-hoc analysis revealed that except for Item 13 favoring Class 2, Items 11 and 12 favored Classes 3 and 1, respectively. Similar to MC items in section 1, low- and moderate-level

examinees show more tendency to get higher scores on MC items because of test-taking strategies and guessing factors. As noted above, males are also more risk averters than females in MC items.

Items 14 and 15 also indicated a significant advantage for Class 3. For these items, examinees should choose two out of five choices. The better performance of Class 3 members can be justified by two main reasons. First, there is an overlap in wording between oral input, the correct choices, and the distractors. Researchers showed that lexical overlap between the text and the answer options impacts item difficulty (Freedle & Kostin, 1996); the more overlap, the easier item will be. However, the values of outfit MNSQ for Item 14 show that the better performance of Class 3 members is unexpected, likely due to the lack of common correct answers in the responses of the group members. Therefore, some examinees in this group might have correctly answered the item by chance.

Second, although the answers are clearly stated in the oral stimuli, it seems that the combined length of these two items and the provided options impacted the performance of examinees. Researchers argued that the presence of two consecutive items in the IELTS listening can increase item difficulty and negatively affect the comprehension of examinees (Coleman & Heap, 1998). These items might be assessing examinees' reading speed and memory span, which are not relevant to the listening ability being measured. Such challenges affected the performance of high-ability examinees in Class 2. Additionally, it appears that examinees in Class 2 missed these two items out of carelessness. This supposition is supported by the outfit MNSQ patterns of this group, suggesting that their performance on these items was abnormal. Overall, the results indicate that the interaction between the combined length of items and the overlap in wording can significantly impact the performance of examinees.

Items 16–20. These items measure examinees' linguistic knowledge, world knowledge, ability to understand detailed or specific factual information, and ability to integrate listening, reading, short-term memory span, and writing abilities. Our post-hoc analysis showed that (1) Item 17 contributed to variability across all the classes, favoring Class 3; (2) Item 16 favored Class 2; (3) Items 17 and 19 functioned in favor of Class 3; and (4) Items 18 and 20 favored Class 1. The results show that some test items seem to have imposed listening-construct-irrelevant challenges on examinees. For instance, a portion of the input including the correct response for Item 17 is "… full cooked breakfast and evening entertainment …". Examinees should simultaneously read the stem (i.e., Full cooked breakfast Entertainment in the …), listen to the input, and write their answer. The item also requires examinees to mentally rearrange the vocabulary. Due to the difficulty of reading this specific item format simultaneously, this rearrangement might place additional memory demands on examinees, particularly because the response to Item 16 is just a few words earlier in the oral input. This indicates that the test item format poses a challenge for examinees (Field, 2009). Such item formats require understanding numerous details, which hinders deep comprehension of the material (Field, 2009). Therefore, this could be a possible reason why high-level examinees (Class 2) missed Item 17 compared to low- and moderate-level examinees, who might have missed Item 16 and only focused on Item 17.

Another noteworthy point that warrants special consideration is the effect of item instruction on the performance of examinees. The instruction of Items 16–20

is "Complete the sentences below. Write NO MORE THAN TWO WORDS AND/OR A NUMBER for each answer." The interpretation of the statement can vary depending on linguistic, cultural, and educational background of examinees. Researchers have argued how language, culture, and education shape interpretation (Hofstede, 2001; Weir, 1990). Therefore, more research is required to illuminate to what extent item instruction can cause different item performance across examinees with different nationalities and L1 background. This variability in the interpretation of the item instruction might have caused a misunderstanding among examinees, although the word limit for all production items is three. The answer to Item 18 is "(four-course) dinner," with "four-course" being optional. It appears that many high-level examinees may have erroneously assumed that providing all three words was necessary for the correct response, and some of them may have written 'four course dinner' instead of 'four-course dinner'. Despite the instructions not clearly stipulating the required word count, it seems that the cognitive load of maintaining three words simultaneously proved challenging for these examinees. Conversely, examinees with low and moderate abilities who managed to recall the final word, 'dinner', were able to successfully include it in their response. It is also possible that due to misspelling, high-level examinees did not receive any partial credit, highlighting the importance of using polytomous scoring scale (Bodie et al., 2011). These findings are totally in line with Aryadoust (2015).

The better performance of examinees with low to moderate abilities diverges with Coleman and Heap (1998) study stating that table completion items are the most challenging items for examinees in the listening section of the IELTS, because test formats requiring examinees to write words in gaps or to compose responses to short-answer questions impose a substantial additional demand unrelated to the construct of listening ability. The variances observed among the classes may also be linked to gender, with females showing a better performance in tasks such as table completion.

Section 3

Items 23. In this section (academic dimension), there is a conversation between three students on campus talking about study programs. The primary dimension targets for these items are linguistic knowledge, world knowledge, ability to make paraphrases, ability to understand detailed or specific factual information, and making inferences. Our post-hoc analysis indicated that the item favored examinees with low to moderate abilities, likely due to test-taking strategies and guessing, with males exhibiting more tendency to take a risk than females.

Items 21–22 and 24–30. The attributes are linguistic knowledge, world knowledge, ability to make paraphrases, ability to understand detailed or specific factual information, and ability to integrate listening, reading, short-term memory span, and writing abilities. The post-hoc analysis showed that Items 27 and 28 favored Class 2, and the remaining items mostly functioned in favor of moderate-level examinees (Class 1), followed by Class 2.

The instruction for the items is "Complete the sentences below. Write NO MORE THAN TWO WORDS AND/OR A NUMBER for each answer." The performance of examinees might have been impacted by the item instructions, with high-level

examinees potentially missing easy items due to carelessness and unclear guidance. This finding disagrees with previous studies (e.g., Aryadoust, 2012) reporting limited production items tend to favor high-ability examinees due to their greater difficulty. However, in this study, low- and moderate-level examinees had a better performance on such items. Additionally, females generally tend to exhibit better performance in SC items than males (Aryadoust, 2012).

Section 4

Items 31–33 and 38–40. In this section (academic dimension), a guest university lecturer presents a talk about a bird of prey (Peregrine Falcons). The primary dimensions are linguistic knowledge, world knowledge, ability to make paraphrases, ability to understand detailed or specific factual information, making inferences, and ability to integrate listening, reading, short-term memory span, and writing abilities. Our post-hoc analysis showed that except for Item 31 favoring Class 3, the other items functioned in favor of Class 2, followed by Class 1. There are two possible reasons for the better performance of low-level examinees in Item 31. First, the necessary information for answering the item is located at the beginning of the oral input, making answering the item easier. Research has shown that the location of required information for answering a test item affects item difficulty, with items located at the beginning of input being easier compared to items located in the middle or at the end (Yanagawa & Green, 2008). Second, the instruction for Items 31–33 is "Complete the sentences below. Write NO MORE THAN THREE WORDS for each answer." Similar to TC items in section two, the performance of moderate- and high-level examinees might have been affected by the lack of clarity in the item instruction. Females also tend to outperform their male counterparts in TC items (Aryadoust, 2012).

Items 34–37. The attributes for these items are linguistic knowledge, world knowledge, ability to make paraphrases, ability to understand detailed or specific factual information, making inferences, and ability to integrate listening, reading, short-term memory span, and writing abilities. The post-hoc analysis showed that except for Item 36, the remaining items functioned in favor of high-level examinees (Class 2). Item responses of examinees indicate that the item appears to have presented challenges to examinees that are unrelated to the listening construct being assessed. Item 36 was designed to assess the comprehension of specific details. The correct answer for the item is "leave the nest". However, some high-level examinees wrote "live" instead of "leave" on their answer sheets, suggesting that phoneme recognition abilities, while different from listening comprehension, played a notable role. It could be a viable reason to state that some high-level examinees missed this item out of carelessness.

Unlike TC items in section 2 where low- and moderate-level examinees outperformed their high-level counterparts, Class 2 members had a better performance on TC items in section 4. This contradiction between the findings can be ascribed to the effect of other factors, such as more complex items, more paraphrased content, longer sentences, more multitasking items, and faster speech delivery in section 4 compared to section 2. Therefore, the location of item formats can affect item difficulty, especially in the IELTS listening test. Females also exhibit better performance in SC items than males (Aryadoust, 2012).

Implications, limitations, and directions for future research

This study holds several methodological, theoretical, and pedagogical implications. From a methodological standpoint, the current study builds upon and extends prior applications of MRM in educational testing, particularly in the identification of latent class DIF and the exploration of multiple profiles in L2 listening comprehension. From a theoretical standpoint, gaining a deeper understanding of individual difference profiles would allow scholars to model language processing in a more cohesive manner and develop more robust theories and models of language acquisition and performance, especially with respect to L2 listening comprehension. From a pedagogical perspective, the findings of this study underscore the significance of detecting latent class DIF and exploring multiple profiles in L2 listening comprehension. Analyzing each profile aids test developers in understanding the test-taking patterns of examinees and the mental processes employed by them to achieve correct responses. It also provides teachers with insights into individual differences among students and their learning status. Consequently, they can adapt their classroom instructions to enhance students' learning, refine instructional materials and activities, customize instruction based on students' needs and challenging areas, and ultimately provide feedback to promote effective teaching and learning.

Several limitations should also be considered when interpreting the results of this study. Firstly, the examinees who took the listening test were drawn from a relatively small international population. Additional research with a larger and more diverse population is required to corroborate the generalizability of the findings. Secondly, this study considered only a limited number of item types (e.g., multiple-choice, map labeling, table completion, and sentence completion) in the administered test. In future studies, researchers may consider incorporating a broader range of item formats such as classifying and matching items in their tests. Third, it is important to note that the sample size for the present study may be considered relatively modest for the application of MRM. One limitation of MRM is its requirement for larger sample sizes, particularly when extending it to polytomous models, where increasing the number of latent classes necessitates larger samples. As argued by von Davier and Rost (1995), to obtain accurate parameter estimates in a multidimensional analysis, the required sample size should be multiplied by the number of classes. In previous studies, sample sizes greatly varied from 99 to 251,278. A number of studies showed that despite the potential for higher standard error of estimates with small sample sizes and an increase in test items and categories, MRM can still produce stable parameter estimates with relatively modest sample sizes (e.g., Aryadoust, 2015; Frick et al., 2015). Future studies can use larger sample sizes to apply the MRM to explore individual difference profiles across language skills and components.

Another limitation of this study was the inability to fully characterize the three latent classes identified due to the lack of access to covariates or variables typically used for such characterization. The study relied solely on item responses from IELTS examinees. Although this practice of analysis is the main purpose of using the MRM, this may constrain the depth of insight into the latent classes' characteristics and associations with relevant covariates, potentially leading to a less nuanced understanding of student performance or behavior. Therefore, an intriguing avenue for further investigation involves considering a range of covariates and/or factors to provide a comprehensive

picture of individual difference profiles in L2 listening comprehension. Covariates such as lexico-grammatical knowledge, metacognitive strategies, age, gender, working memory capacity, self-efficacy, motivation, and contextual factors could provide valuable insights into the analysis of listening profiles.

In particular, it is important to consider working memory capacity in future studies because it significantly affects the performance of examinees with different working memory levels. Working memory capacity influences how well individuals can process and retain information, which is crucial in tasks such as L2 listening comprehension (Goh & Vandergrift, 2022). By accounting for this variable, researchers can better understand and interpret the differences in performance among examinees. Also, the results showed that there should be an interaction among covariates (e.g., gender and item types) causing DIF. As a recent development in DIF analysis, future studies can apply advanced tree models (e.g., Grassi & Tarantino, 2023; Henninger et al., 2023) for investigating DIF of the IELTS listening test.

Moreover, the assumptions about what different parts of the IELTS listening test measure, as shown in Table 1 and argued in the discussion, were based on the subjective judgment of the authors of this study and those in similar studies (e.g., Aryadoust, 2012; Buck, 2001). These descriptions are broad and refer to attributes collectively measured by each item type across the four parts of the test. For instance, MC items are intended to measure a wide range of traits, such as the ability to attentively listen and comprehend the main idea and details of oral stimuli, identify key information like main points and supporting details, understand context, use inference skills to deduce implied meanings, and distinguish between similar options to eliminate distractors. However, a broad description like this does not specify which attributes are measured by each MC item. Typically, an item cannot measure the entire range of attributes mentioned. Future studies could employ empirical methods, such as diagnostic classification models (DCMs; Kunina-Habenicht et al., 2009; Ravand & Baghaei, 2020; Ravand et al., 2019), to identify the attributes measured by each individual item rather than by each item type as a whole. This approach would provide a clearer understanding of the causes of DIF when an item is flagged for DIF. In the absence of covariates to explain the differences in the latent classes, one can refer to the abilities and traits measured by different items on the test to explain the differential performances of the three latent classes.

The *TAM* package (Robitzsch et al., 2024) employed in this study uses equal-mean-difficulty anchor method, which assumes the mean of difficulty across items is equal among classes or (sub)groups (commonly equal to zero). Therefore, in this study, the mean of difficulty parameters across items was constrained to be equal to zero for each latent class in MRM analysis. The assumption might be violated when the direction of DIF is unbalanced (Kopf et al., 2015) between classes. Another popular anchor method is the constant anchor method, where a set of items as DIF-free items are prespecified (Meade & Wright, 2012). Several researchers have shown that the use of constant anchor method in mixture IRT models for identifying latent classes leads to better model-data fit (Chen et al., 2023), and that the purification for the constant anchor method reduces the type I error rate in DIF analysis (Meade & Wright, 2012). Future studies can thus use the constant anchor method for exploring multiple profiles of L2 listeners.

Numerous researchers have developed both unidimensional (e.g., Sen et al., 2019; Tseng & Wang, 2021) and multidimensional generalizations (e.g., von Davier, 2008) of MRM, demonstrating successful application of the models to language data. Researchers

could further extend the application of MRM using these models to investigate their effectiveness in exploring profiles of examinees within the domain of educational measurement and language assessment.

Of particular interest are also potential applications for MRM in the area of PLP listening tests. Previous studies have predominantly utilized WLP tests to explore multiple profiles of listeners and investigate latent class DIF. However, no prior research has ventured into applying the MRM to PLP tests to analyze cognitive processes of examinees and identify their solution patterns while answering a set of test items. Future studies could delve into examining to what extent the introduction of visual input into listening tests may alter listening processing patterns. A number of researchers have advocated for the inclusion of nonverbal information such as gestures, posture, facial expressions, and body movement, often observed in authentic communications, as integral components of the L2 proficiency construct (Lesnov, 2022; Park et al., 2022; Wagner, 2013; Wolvin & Coakley, 1993).

Finally, future studies could directly hypothesize and investigate the bottom-up and top-down listening processing strategies employed by examinees, which yield valuable insights into language comprehension and test performance. Various methodologies such as eye-tracking technology, cognitive interviews, or neuroimaging techniques can be used to observe and analyze how examinees engage in these processes. By understanding which strategies are more effective for different individuals or in various contexts, educators and test developers can tailor instructional approaches and assessment designs to better support language learning and assessment.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Research data policy and data availability statements

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable requests.

Funding

The author(s) received no specific funding for this work from any funding agencies.

ORCID

Farshad Effatpanah http://orcid.org/0000-0003-3970-5588
Purya Baghaei http://orcid.org/0000-0002-5765-0413
Hamdollah Ravand http://orcid.org/0000-0002-8757-3850
Olga Kunina-Habenicht http://orcid.org/0000-0002-1646-8260

References

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, *22*(3), 37–51. https://doi.org/10.1111/j.1745-3992.2003.tb00136.x

Alavi, S. M., Kaivanpanah, S., & Masjedlou, A. P. (2018). Validity of the listening module of international English language testing system: Multiple sources of evidence. *Language Testing in Asia*, *8*(1), 1–17. https://doi.org/10.1186/s40468-018-0057-4

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. https://doi.org/10.1111/j.1745-3984.2011.00146.x

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

Aryadoust, V. (2012). Differential item functioning in while-listening performance tests: The case of IELTS listening test. *International Journal of Listening*, *26*(1), 40–60. https://doi.org/10.1080/10904018.2012.639649

Aryadoust, V. (2015). Fitting a mixture Rasch model to EFL listening tests: The role of cognitive and background variables in explaining latent differential item functioning. *International Journal of Testing*, *15*(3), 216–238. https://doi.org/10.1080/15305058.2015.1004409

Aryadoust, V, T (2018). Taxonomies of listening skills. In J. I. Liontas, & M. DelliCarpini (Eds.), *The TESOL encyclopedia of English language teaching*. (pp. 1–8). https://doi.org/10.1002/9781118784235.eelt0577

Aryadoust, V., Goh, C. C. M., & Kim, L. O. (2011). An investigation of differential item functioning in the MELAB listening test. *Language Assessment Quarterly*, *8*(4), 361–385. https://doi.org/10.1080/15434303.2011.628632

Aryadoust, V., Min, S., & Chen, X. (2024). Investigating differential item functioning across interaction variables in listening comprehension assessment. *Studies in Educational Evaluation*, *80*, 101322. https://doi.org/10.1016/j.stueduc.2024.101322

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. https://doi.org/10.7275/n191-pt86

Baghaei, P., Kemper, C. J., Reichert, M., & Greiff, S. (2019). Applying the mixed Rasch model in assessing reading comprehension. In V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment Volume II: Advanced methods*. (pp. 15–32) Routledge.

Banerjee, J., & Papageorgiou, S. (2016). What's in a topic? Exploring the interaction between test-taker age and item content in high-stakes testing. *International Journal of Listening*, *30*(1-2), 8–24. https://doi.org/10.1080/10904018.2015.1056876

Bejar, I., Douglas, D., Jamieson, J., Nissan, S., & Turner, J. (2000). *TOEFL 2000 listening framework: A working paper*. (TOEFL Monograph Series No. MS-19) Educational Testing Service.

Bodie, G. D., & Worthington, D. L. (2010). Revisiting the listening styles profile (LSP-16): A confirmatory factor analytic approach to scale validation and reliability estimation. *International Journal of Listening*, *24*(2), 69–88. https://doi.org/10.1080/10904011003744516

Bodie, G. D., Worthington, D., & Fitch-Hauser, M. (2011). A comparison of four measurement models for the Watson-Barker Listening Test (WBLT)-Form C. *Communication Research Reports*, *28*(1), 32–42. https://doi.org/10.1080/08824096.2011.540547

Bodie, G. D., Worthington, D. L., & Gearhart, C. C. (2013). The listening styles profile revised (LSP-R): A scale revision and evidence for validity. *Communication Quarterly*, *61*(1), 72–90. https://doi.org/10.1080/01463373.2012.720343

Bodie, G. D., Winter, J., Dupuis, D., & Tompkins, T. (2020). The echo listening profile: Initial validity evidence for a measure of four listening habits. *International Journal of Listening*, *34*(3), 131–155. https://doi.org/10.1080/10904018.2019.1611433

Bolger, N., & Kellaghan, T. (1990). Method of measurement and gender differences in scholastic achievement. *Journal of Educational Measurement*, *27*(2), 165–174. https://doi.org/10.1111/j.1745-3984.1990.tb00740.x

Bond, T. G., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences*. (4th Ed.) Routledge.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Bourdeaud'Hui, H., Aesaert, K., & van Braak, J. (2021). Exploring the validity of a comprehensive listening test to identify differences in primary school students' listening skills. *Language Assessment Quarterly*, *18*(3), 228–252. https://doi.org/10.1080/15434303.2020.1860059

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

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing*, *15*(2), 119–157. https://doi.org/10.1191/026553298667688289

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. (2nd Ed.) Springer. https://doi.org/10.1007/b97636

Carlton, S. T., & Harris, A. M. (1992). Characteristics associated with differential item functioning on the scholastic aptitude test: Gender and majority/minority group comparisons. *ETS Research Report Series*, *1992*(2), i–143. https://doi.org/10.1002/j.2333-8504.1992.tb01495.x

Chang, A. C.-S., & Read, J. (2013). Investigating the effects of multiple-choice listening test items in the oral versus written mode on L2 listeners' performance and perceptions. *System*, *41*(3), 575–586. https://doi.org/10.1016/j.system.2013.06.001

Chen, H., & Chen, J. (2021). Investigating the relationships between listening skills and genre competence through cognitive diagnosis approach. *Sage Open*, *11*(4), 1–14. https://doi.org/10.1177/21582440211061342

Chen, C., W., Andersson, B., & Zhu, J. (2023). A factor mixture model for item responses and certainty of response indices to identify student knowledge profiles. *Journal of Educational Measurement*, *60*(1), 28–51. https://doi.org/10.1111/jedm.12344

Choi, I. H., Paek, I., & Cho, S. J. (2017). The impact of various class-distinction features on model selection in the mixture Rasch model. *The Journal of Experimental Education*, *85*(3), 411–424. https://doi.org/10.1080/00220973.2016.1250208

Chon, Y. V., & Shin, T. (2019). Profile of second language learners' metacognitive awareness and academic motivation for successful listening: A latent class analysis. *Learning and Individual Differences*, *70*, 62–75. https://doi.org/10.1016/j.lindif.2019.01.007

Cid, J., Wei, Y., Kim, S., & Hauck, C. (2017). *Statistical analyses for the updated TOEIC® listening and reading test (Research Memorandum No. RM-17-05)*. Educational Testing Service.

Cohen, A., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, *42*(2), 133–148. https://doi.org/10.1111/j.1745-3984.2005.00007

Cole, N. S. (1997). *The ETS gender study: How males and females perform in educational settings*. Educational Testing Service, College Entrance Examination.

Coleman, G., Heap, S. (1998). *The misinterpretation of directions for the questions in the Academic Reading and Listening sub-tests of the IELTS test*. (Research Report No. 1, IELTS Australia). URL: https://ielts.org/researchers/our-research/research-reports/the-misinterpretation-of-directions-for-the-questions-in-the-academic-reading-and-listening-sub-tests-of-the-ielts-test

Curley, W., & Schmitt, A. P. (1993). Revising SAT®-Verbal items to eliminate differential item functioning. *ETS Research Report Series*, *1993*(2), i–18. https://doi.org/10.1002/j.2333-8504.1993.tb01572.x

De Ayala, R. J., & Santiago, S. Y. (2017). An introduction to mixture item response theory models. *Journal of School Psychology*, *60*, 25–40. https://doi.org/10.1016/j.jsp.2016.01.002

Du, G., & Man, D. (2022). Person factors and strategic processing in L2 listening comprehension: Examining the role of vocabulary size, metacognitive knowledge, self efficacy, and strategy use. *System*, *107*, 102801. https://doi.org/10.1016/j.system.2022.102801

Effatpanah, F. (2019). Application of cognitive diagnostic models to the listening section of the International English Language Testing System (IELTS). *International Journal of Language Testing*, *9*(1), 1–28. https://www.ijlt.ir/article_114295.html

Effatpanah, F., Baghaei, P., & Karimi, M. N. (2024). A mixed Rasch model analysis of multiple profiles in L2 writing. *Assessing Writing*, *59*, 100803. https://doi.org/10.1016/j.asw.2023.100803

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates Publishers.

Estaji, M., & Banitalebi, Z. (2023). A study of test-taking strategies of Iranian IELTS repeaters: Any change in the strategy use? *International Journal of Testing*, *23*(3), 205–230. https://doi.org/10.1080/15305058.2023.2195662

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) Pty Ltd & British Council.

Field, J. (2013). Cognitive validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening.* (pp. 77–151) Cambridge University Press.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement*, *29*(4), 278–295. https://doi.org/10.1177/0146621605275728

Freedle, R., & Kostin, I. (1996). The prediction of TOEFL listening comprehension item difficulty for mini-talk passages: Implications for construct validity. *ETS Research Report Series*, *1996*(2), i–61. https://doi.org/10.1002/j.2333-8504.1996.tb01707.x

Frick, H., Strobl, C., & Zeileis, A. (2015). Rasch mixture models for DIF detection: A comparison of old and new score specifications. *Educational and Psychological Measurement*, *75*(2), 208–234. https://doi.org/10.1177/0013164414536183

Geranpayeh, A., & Kunnan, A. J. (2007). Differential item functioning in terms of age in the certificate in advanced English examination. *Language Assessment Quarterly*, *4*(2), 190–222. https://doi.org/10.1080/15434300701375758

Geranpayeh, A., & Taylor, L. (2008). Examining listening: Developments and issues in assessing second language listening. *Cambridge Research Notes*, *32*, 3–5. https://www.cambridgeenglish.org/images/23151-research-notes-32.pdf

Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications.* (pp. 69–95) Springer. https://doi.org/10.1007/978-1-4612-4230-7_5

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. https://michiganassessment.org/wp-content/uploads/2020/02/20.02.pdf.Res_.InvestigatingtheConstructValidityoftheMELABListeningTestthroughtheRaschAnalysisandCorrelatedUniquenessModeling.pdf

Goh, C. C. M., & Vandergrift, L. (2022). *Teaching and learning second language listening: Metacognition in action.* (2nd Ed.) Routledge.

Graham, S. (2017). Research into practice: Listening strategies in an instructed classroom setting. *Language Teaching*, *50*(1), 107–119. https://doi.org/10.1017/S0261444816000306

Grassi, M., & Tarantino, B. (2023). SEMtree: Tree-based structure learning methods with structural equation models. *Bioinformatics*, *39*(6), 1–9. https://doi.org/10.1093/bioinformatics/btad377

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

Harding, L., Alderson, J. C., & Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles. *Language Testing*, *32*(3), 317–336. https://doi.org/10.1177/0265532214564505

He, L., & Jiang, Z. (2020). Assessing second language listening over the past twenty years: A review within the socio-cognitive framework. *Frontiers in Psychology*, *11*, 2123. https://doi.org/10.3389/fpsyg.2020.02123

Henninger, M., Debelak, R., & Strobl, C. (2023). A new stopping criterion for Rasch trees based on the Mantel–Haenszel effect size measure for differential item functioning. *Educational and Psychological Measurement*, *83*(1), 181–212. https://doi.org/10.1177/00131644221077135

Hickendorff, M., Edelsbrunner, P. A., McMullen, J., Schneider, M., & Trezise, K. (2018). Informative tools for characterizing individual differences in learning: Latent class, latent profile, and latent transition analysis. *Learning and Individual Differences*, *66*, 4–15. https://doi.org/10.1016/j.lindif.2017.11.001

Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations.* (2nd Edition) Sage.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity.* (pp. 129–145) LEA.

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

Humphry, S., & Montuoro, P. (2021). The Rasch model cannot reveal systematic differential Item functioning in single tests: Subset DIF analysis as an alternative methodology. *Frontiers in Education*, *6*, 742560. https://doi.org/10.3389/feduc.2021.742560

Imhof, M., & Janusik, L. A. (2006). Development and validation of the Imhof-Janusik listening concepts inventory to measure listening conceptualization differences between cultures. *Journal of Intercultural Communication Research*, *35*(2), 79–98. https://doi.org/10.1080/17475750600909246

Isbell, D. R., & Kremmel, B. (2020). Test review: Current options in at-home language proficiency tests for making high-stakes decisions. *Language Testing*, *37*(4), 600–619. https://doi.org/10.1177/0265532220943483

Kahneman, D. (1973). *Attention and effort*. Prentice-Hall.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.

Kopf, J., Zeileis, A., & Strobl, C. (2015). Anchor selection strategies for DIF analysis: Review, assessment, and new approaches. *Educational and Psychological Measurement*, *75*(1), 22–56. https://doi.org/10.1177/0013164414529792

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2009). A practical illustration of multidimensional diagnostic skills profiling: Comparing results from confirmatory factor analysis and diagnostic classification models. *Studies in Educational Evaluation*, *35*(2–3), 64–70. https://doi.org/10.1016/j.stueduc.2009.10.003

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Houghton Mifflin.

Lesnov, R. O. (2022). Furthering the argument for visually inclusive L2 academic listening tests: The role of content-rich videos. *Studies in Educational Evaluation*, *72*, 101087. https://doi.org/10.1016/j.stueduc.2021.101087

Li, F., Cohen, A. S., Kim, S.-H., & Cho, S.-J. (2009). Model selection methods for mixture dichotmous IRT models. *Applied Psychological Measurement*, *33*(5), 353–373. https://doi.org/10.1177/0146621608326422

Liao, L., & Yao, D. (2021). Grade-related differential item functioning in general English proficiency test-kids listening. *Frontiers in Psychology*, *12*, 767244. https://doi.org/10.3389/fpsyg.2021.767244

Lin, J., & Wu, F. (2003, April 22-24). *Differential performance by gender in foreign language testing* [Poster presentation]. The Annual Meeting of the National Council on Measurement in Education (NCME), Chicago, IL, U.S.A. URL: https://files.eric.ed.gov/fulltext/ED478206.pdf

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*(2), 878. https://www.rasch.org/rmt/rmt162f.htm

Linacre, J. M. (2024). *A user's guide to WINSTEPS*. Winsteps.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erbaum Associates.

Marx, A., Heppt, B., & Henschel, S. (2017). Listening comprehension of academic and everyday language in first language and second language students. *Applied Psycholinguistics*, *38*(3), 571–600. https://doi.org/10.1017/S0142716416000333

Mazzeo, J., Schmitt, A. P., & Bleistein, C. A. (1993). Sex-related performance differences on constructed-response and multiple-choice sections of Advanced Placement Examinations. *ETS Research Report Series*, *1*, i–29. https://doi.org/10.1002/j.2333-8504.1993.tb01516.x

Meade, A. W., & Wright, N. A. (2012). Solving the measurement invariance anchor item problem in item response theory. *The Journal of Applied Psychology*, *97*(5), 1016–1031. https://doi.org/10.1037/a0027934

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, *58*(4), 525–543. https://doi.org/10.1007/BF02294825

Nishizawa, H. (2023). Construct validity and fairness of an operational listening test with world Englishes. *Language Testing*, *40*(3), 493–520. https://doi.org/10.1177/02655322221137869

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(4), 535–569. https://doi.org/10.1080/10705510701575396

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

O'Neill, K. A., & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In Holland, P. W. & Wainer, H. (Eds.), *Differential item functioning*., (pp. 255–276) Lawrence Earlbaum.

Oshima, T. C., Raju, N. S., & Flowers, C. P. (1997). Development and demonstration of multidimensional IRT-based internal measures of differential functioning of items and tests. *Journal of Educational Measurement*, *34*(3), 253–272. https://doi.org/10.1111/j.1745-3984.1997.tb00518.x

Pae, T.-I. (2004). DIF for examinees with different academic backgrounds. *Language Testing*, *21*(1), 53–73. https://doi.org/10.1191/0265532204lt274oa

Pae, T.-I. (2012). Causes of gender DIF on an EFL language test: A multiple-data analysis over nine years. *Language Testing*, *29*(4), 533–554. https://doi.org/10.1177/0265532211434027

Park, G. P. (2008). Differential item functioning on an English listening test across gender. *TESOL Quarterly*, *42*(1), 115–123. https://doi.org/10.1002/j.1545-7249.2008.tb00212.x

Park, Y., Lee, S., & Shin, S. Y. (2022). Developing a local academic English listening test using authentic unscripted audio-visual texts. *Language Testing*, *39*(3), 401–424. https://doi.org/10.1177/02655322221076024

Preinerstorfer, D., & Formann, A. K. (2012). Parameter recovery and model selection in mixed Rasch models. *British Journal of Mathematical and Statistical Psychology*, *65*(2), 251–262. https://doi.org/10.1111/j.2044-8317.2011.02020.x

R Core Team (2024). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria. URL:https://www.R-project.org

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*(4), 495–502. https://doi.org/10.1007/BF02294403

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, *14*(2), 197–207. https://doi.org/10.1177/014662169001400208

Raquel, M. (2019). The Rasch measurement approach to differential item functioning (DIF) analysis in language assessment research. In V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment (volume I): Fundamental techniques.* (pp. 103–131) Routledge.

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

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. In M. Blegvad (Ed.), *The Danish yearbook of philosophy*. Munksgaard. https://doi.org/10.1163/24689300-01401006

Ravand, H. (2015). Assessing testlet effect, impact, differential testlet, and item functioning using cross-classified multilevel measurement modeling. *Sage Open*, *5*(2), 1–9. https://doi.org/10.1177/2158244015585607

Ravand, H. (2024). Assessing measurement invariance in a university entrance exam: A comparison of multigroup confirmatory factor analysis alignment method vs. multigroup item response theory. *Educational Methods & Psychometrics*, *2*, 11. https://doi.org/10.61186/emp.2024.4

Ravand, H., & Baghaei, P. (2020). Diagnostic classification models: Recent developments, practical issues, and prospects. *International Journal of Testing*, *20*(1), 24–56. https://doi.org/10.1080/15305058.2019.1588278

Ravand, H., Rohani, G., & Firoozi, T. (2019). Investigating gender and major DIF in the Iranian National University Entrance Exam using multiple-indicators multiple-causes structural equation modelling. *Issues in Language Teaching*, *8*(1), 33–61. https://doi.org/10.22054/ilt.2020.49509.460

Ravand, H., Baghaei, P., & Doebler, P. (2019). Examining parameter invariance in a general diagnostic classification model. *Frontiers in Psychology*, *10*, 2930. https://doi.org/10.3389/fpsyg.2019.02930

Richards, J. C. (1983). Listening comprehension: Approach, design, procedure. *TESOL Quarterly*, *17*(2), 219–240. https://doi.org/10.2307/3586651

Robitzsch, A., Kiefer, T., & Wu, M. (2024). *TAM: Test Analysis Modules*. R package version 4.2-21. URL: https://cran.r-project.org/web/packages/TAM

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*(3), 271–282. https://doi.org/10.1177/014662169001400305

Rost, M. (2016). *Teaching and researching listening.* (3rd Ed.) Longman.
Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement*, *20*(4), 355–371. https://doi.org/10.1177/014662169602000404
Rukthong, A., & Brunfaut, T. (2020). Is anybody listening? The nature of second language listening in integrated listening-to-summarize tasks. *Language Testing*, *37*(1), 31–53. https://doi.org/10.1177/0265532219871470
Scheuneman, J. D., & Gerritz, K. (1990). Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. *Journal of Educational Measurement*, *27*(2), 109–131. https://doi.org/10.1111/j.1745-3984.1990.tb00737.x
Sen, S. (2018). Spurious latent class problem in the mixed Rasch model: A comparison of three maximum likelihood estimation methods under different ability distributions. *International Journal of Testing*, *18*(1), 71–100. https://doi.org/10.1080/15305058.2017.1312408
Sen, S., & Cohen, A. (2019). Applications of mixture IRT models: A literature review. *Measurement: Interdisciplinary Research and Perspectives*, *17*(4), 177–191. https://doi.org/10.1080/15366367.2019.1583506
Sen, S., Cohen, A. S., & Kim, S. H. (2019). Model selection for multilevel mixture Rasch models. *Applied Psychological Measurement*, *43*(4), 272–289. https://doi.org/10.1177/0146621618779990
Seo, D., Taherbhai, H., & Frantz, R. (2016). Psychometric evaluation and discussions of English language learners' listening comprehension. *International Journal of Listening*, *30*(1-2), 47–66. https://doi.org/10.1080/10904018.2015.1065747
Shin, S. Y., Lee, S., & Lidster, R. (2021). Examining the effects of different English speech varieties on an L2 academic listening comprehension test at the item level. *Language Testing*, *38*(4), 580–601. https://doi.org/10.1177/0265532220985432
Shohamy, E., & Inbar, O. (1991). Validation of listening comprehension tests: The effect of text and question type. *Language Testing*, *8*(1), 23–40. https://doi.org/10.1177/026553229100800103
Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. *Psychological Methods*, *11*(4), 402–415. https://doi.org/10.1037/1082-989X.11.4.402
Swaminathan, H., & Rogers, H. J. (2000). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361–370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x
Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4
Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics.* (6th Ed.) Pearson Education.
Tay, L., Newman, D. A., & Vermunt, J. K. (2011). Using mixed-measurement item response theory with covariates (MM-IRT-C) to ascertain observed and unobserved measurement equivalence. *Organizational Research Methods*, *14*(1), 147–176. https://doi.org/10.1177/1094428110366037
Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning.* (pp. 67–113) Erlbaum.
Tseng, M. C., & Wang, W. C. (2021). The Q-Matrix anchored mixture Rasch model. *Frontiers in Psychology*, *12*, 564976. https://doi.org/10.3389/fpsyg.2021.564976
Van Nijlen, D., & Janssen, R. (2011). Measuring mastery across grades: An application to spelling ability. *Applied Measurement in Education*, *24*(4), 367–387. https://doi.org/10.1080/08957347.2011.607064
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) Information Age Publishing.
von Davier, M., & Rost, J. (1995). Polytomous mixed Rasch models. In G. H. Fischer & I. W. Molennar (Eds.), *Rasch models: Foundations, recent developments, and applications.* (pp. 371–379) Springer Verlag.
Wagner, E. (2013). An investigation of how the channel of input and access to test questions affect L2 listening test performance. *Language Assessment Quarterly*, *10*(2), 178–195. https://doi.org/10.1080/15434303.2013.769552

Wang, W.-C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *The Journal of Experimental Education*, *72*(3), 221–261. https://doi.org/10.3200/JEXE.72.3.221-261

Wang, W., Liu, Y., & Liu, H. (2022). Testing differential item functioning without predefined anchor items using robust regression. *Journal of Educational and Behavioral Statistics*, *47*(6), 666–692. https://doi.org/10.3102/10769986221109208

Watson, K. W., Barker, L. L., & Weaver, J. B. (1995). The listening styles profile (LSP-16): Development and validation of an instrument to assess four listening styles. *International Journal of Listening*, *9*(1), 1–13. https://doi.org/10.1080/10904018.1995.10499138

Weir, C. (1990). *Communicative language testing*. Prentice Hall.

Willingham, W. W., & Cole, N. S. (1997). Fairness issues in test design and use. In Willingham, W.W. & Cole, N. S. (Eds.), *gender and fair assessment.* (pp. 227–346) Lawrence Erlbaum. https://doi.org/10.4324/9781315045115

Wolvin, A. D. (2013). Understanding the listening process: Rethinking the "one size fits all" model. *International Journal of Listening*, *27*(2), 104–106. https://doi.org/10.1080/10904018.2013.783351

Wolvin, A. D., & Coakley, C. G. (1993). A listening taxonomy. In A. D. Wolvin & C. G. Coakley (Eds.), *Perspectives on listening.* (pp. 15–22) Ablex.

Yanagawa, K., & Green, A. (2008). To show or not to show: The effects of items stems and answer options on performance on a multiple-choice comprehension test. *System*, *36*(1), 107–122. https://doi.org/10.1016/j.system.2007.12.003

Yuan, K. H., Liu, H., & Han, Y. (2021). Differential item functioning analysis without a priori information on anchor items: QQ plots and graphical test. *Psychometrika*, *86*(2), 345–377. https://doi.org/10.1007/s11336-021-09746-5

Zansen, A. V., Hilden, R., & Laihanen, E. (2022). The multimodal listening test in a high stakes context: Gender-neutral or not? *International Journal of Listening*, *36*(2), 152–170. https://doi.org/10.1080/10904018.2021.1993446

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*(2), 223–233. https://doi.org/10.1080/15434300701375832