Iranian Journal of Applied Language Studies
University of Sistan and Baluchestan
Uppsala University
Print ISSN: 2008-5494 Online ISSN: 2322-3650
Homepage: https://ijals.usb.ac.ir

Examining Local Item Dependence in the IELTS Academic Listening Module: A Rasch Testlet Model Approach

Navid Reza Zolfaghari, Purya Baghaei, Zohreh Seifoori

Ph.D. Candidate, Department of English, Islamic Azad University, Tehran, Iran, Email: navidr.zolfaghari@iau.ac.ir

Associate Professor, Department of English, Islamic Azad University, Mashhad, Iran, Email:baghaei.purya@gmail.com

Associate Professor, Department of English, Islamic Azad University, Tehran, Iran, Email: zseifoori@iau.ac.ir

Abstract

This study investigates the presence and magnitude of local item dependence (LID) in the IELTS Academic Listening module through the application of the Rasch Testlet Model (RTM). While the unidimensional Rasch Model (RM) assumes item independence, the RTM accounts for shared variance among items nested within common testlets. Using data from forty listening items organized into four testlets, the study compares the fit and reliability indices of the RM and RTM. The findings indicate that although all items fit the Rasch Model based on infit and outfit statistics, notable testlet effects emerged, particularly in Testlets 1 and 2. The RTM exhibited superior model fit indices (lower -2LL, AIC, CAIC, and BIC values) compared to the RM, suggesting that accounting for LID yields a more accurate representation of the data. However, the magnitude of LID in the listening module was considerably smaller than in the reading module, likely due to the transient nature of listening input, which limits excessive engagement with text and item interdependence. These results underscore the importance of addressing testlet effects in high-stakes listening assessments to enhance measurement precision and construct validity.

Keywords: IELTS listening, local item dependence, Rasch Testlet Model, Testlet effect, test validation

1. Introduction

Listening comprehension is broadly acknowledged as one of the most cognitively challenging components of second language (L2) proficiency evaluation, necessitating the swift integration of bottom-up decoding techniques with top-down interpretative approaches under time limitations (Buck, 2001; Field, 2008; Goh & Vandergrift, 2022). In high-pressure standardized tests like the International English Language Testing System (IELTS), the listening component is essential for assessing candidates' receptive abilities. Nonetheless, the accuracy of the conclusions made from these tests relies not only on the representativeness of the content but also on the test's psychometric features, especially the premise of local item independence (Bond & Fox, 2015; Hambleton et al., 1991).

In test design, local item dependence manifests when there are correlations among responses to items that exceed what would be anticipated based on the assessed latent ability (Sireci et al., 1991; Yen, 1984, 1993). A frequent origin of LID is the use of testlets. The testlet refers to as collections of items linked to a common stimulus, like an audio clip in listening assessments or a text excerpt in comprehension evaluations (Bradlow et al., 1999; Wainer & Kiely, 1987; Wainer & Lewis, 1990). The IELTS Academic Listening section, for instance, includes 40 questions structured around four distinct audio clips. Although this framework improves credibility and contextual consistency, it also brings forth possible item interrelatedness (IELTS, n.d.; British Council, n.d.; Cambridge English, n.d.).

The Rasch Model posits that each item independently contributes to the assessment of the latent trait (Bond & Fox, 2015; Wright et al, 1994). The presence of LID can lead to biased parameter estimates, distorted item fit statistics, and artificially inflated reliability (Hambleton et al., 1991; Marais & Andrich, 2008; Thissen et al., 1989). Consequently, investigators have suggested advanced models, like the Rasch Testlet Model, to address within-testlet dependencies by incorporating testlet-specific random effects (DeMars, 2006; Wainer et al., 2007Wang & Wilson, 2005). These models distinguish the actual latent trait variance from the residual variance resulting from common stimuli, providing more precise estimates of individual ability and item difficulty (Baghaei & Aryadoust, 2014; Rijmen, 2009).

The topic of testlet effects has garnered more attention in recent years, especially in the realms of large-scale language assessment and global evaluation settings (e.g., PISA, TOEFL, and Cambridge English tests). Research indicates that neglecting LID may result in an inflated perception of reliability and skewed conclusions regarding item difficulty or discrimination (DeMars, 2006; Sireci et al., 1991; Wainer & Wang, 2001). In contrast, utilizing RTM or bifactor models typically leads to enhanced model fit indices—such as reduced deviance, AIC, BIC, and CAIC values—suggesting that incorporating testlet structure more accurately represents the underlying data (Lim, 2024; Rijmen, 2009; Wainer et al., 2007).

Although testlet effects have been thoroughly researched in reading comprehension, there has been less exploration of their influence on listening evaluations, where the stimuli are fleeting and not visually available (Baghaei & Aryadoust, 2014; Bradlow et al., 1999). Real-time responses are given to listening test items as recordings are played, which might constrain extensive cognitive interaction among items and consequently lessen the extent of LID. Nevertheless, studies indicate that even small testlet effects can impact score accuracy and the validity of the test (Marais & Andrich, 2008; Wang & Wilson, 2005; Zhu et al., 2019). In situations such as IELTS, where scores affect academic placements and career prospects, grasping and modeling these relationships is crucial for maintaining fairness and construct validity (Weir & O'Sullivan, 2011).

Recent empirical studies have utilized RTM to explore testlet effects in both L1 and L2 environments. For instance, DeMars (2006) showed that considering testlet variance in reading comprehension assessments lowered reliability estimates to more accurate levels. Baghaei and Aryadoust (2014) discovered that the RTM offered a superior fit compared to the conventional Rasch model for evaluating local dependence resulting from shared stimuli in language assessment. Likewise, Lim (2024) stressed that LID ought to be regularly assessed in validation research of testlet-based evaluations. In a different effort, Baghaei et al. (2025) examined testlet effects in the PIRLS and ePIRLS 2016 assessment, showing that model fit indices consistently supported the Rasch testlet model rather than the unidimensional model. These results lead to the conclusion that overlooking testlet structure might mask essential elements of measurement error and validity evidence.

Under the IELTS framework, earlier studies have primarily concentrated on the reading module, uncovering significant testlet effects and notable local dependencies between items associated with the same passage (Aryadoust & Zhang, 2016). Conversely, the listening module has gained minimal focus, even though it shares a similar framework and comparable psychometric principles. Considering that the auditory modality presents distinct cognitive and perceptual difficulties—like memory demand and restricted chances for stimulus review (Field, 2008; Goh & Vandergrift, 2022)—it is reasonable to assume that the level of LID in listening varies both qualitatively and quantitatively from that in reading.

The present study therefore aims to examine the extent of testlet effects in the IELTS Academic Listening module and to compare the model fit and reliability indices between the unidimensional Rasch Model and the Rasch Testlet Model. Specifically, it addresses the following research questions:

1. To what extent does the IELTS Academic Listening module exhibit testlet effects?
2. Does the Rasch Testlet Model (when LID is accounted for) fit significantly better than the unidimensional Rasch Model (when LID is ignored) in the IELTS Listening section?

Through empirical comparisons of the fit statistics and reliability estimates from each model, this research aids in the psychometric validation of IELTS Listening and offers evidence-based insights for enhancing score interpretation. It also broadens the literature on LID in language

testing by demonstrating how the RTM can be utilized in auditory comprehension assessment—an aspect that is still underexamined despite its importance to communicative competence in academic and professional settings.

2. Literature Review

Listening comprehension is widely conceptualized as a complex, time-bound process that requires rapid coordination of bottom-up decoding (e.g., phonological processing, segmentation, lexical access) and top-down meaning construction (e.g., inferencing, schematic knowledge, pragmatic interpretation), often under conditions of limited control over input (Buck, 2001; Field, 2008; Goh & Vandergrift, 2022). In high-stakes standardized tests, listening section is typically operationalized through short to medium-length audio segments followed by items that target comprehension. These design aim to approximate real-world listening demands; however, they also introduce measurement challenges tied to transient input, memory load, and the structured bundling of items around shared stimuli (Buck, 2001; Field, 2008; Goh & Vandergrift, 2022).

In the field of language testing validity, defining what a test measures and interpreting its scores require support from various types of evidence. This includes evidence related to psychometric functioning, cognitive processing, and consequences of use (AERA, APA, & NCME, 2014; Chapelle et al., 2008; Kane, 2006, 2013; Weir, 2005). For listening assessments, a recurring concern is whether scores represent listening ability as intended, rather than method effects introduced by test design features such as item clustering, item formats, or shared passage properties. This concern becomes especially salient in assessments like IELTS, where listening scores have material implications for educational and professional opportunities (Kane, 2013; Weir & O'Sullivan, 2011).

A foundational assumption of classical and modern measurement models is that item responses are conditionally independent given the latent trait (Hambleton et al., 1991). In IRT terms, local item dependence occurs when responses to items remain correlated after conditioning on ability, often due to shared content, format, strategies, or common stimuli (Thissen et al., 1989; Yen, 1984, 1993). LID is not merely a technical nuisance: it can inflate test information, distort standard errors, bias item parameter estimates, and generate overly optimistic reliability—ultimately undermining the validity of interpretations based on total scores (DeMars, 2006; Hagell, 2026; Marais & Andrich, 2008; Sireci et al., 1991;).

Empirical and methodological research has shown that LID may remain undetected if analysts rely only on conventional item-level fit statistics (e.g., infit/outfit under Rasch), because items can appear to fit well even when residual correlations within stimulus-linked clusters are present (Marais & Andrich, 2008; Sireci et al., 1991; Yen, 1984). This phenomenon matters particularly for testlet-based assessments, where items are intentionally grouped around a shared stimulus, making some degree of dependence plausible by design rather than by flaw.

The concept of a testlet—a bundle of items linked to a common stimulus—was introduced to describe clustered item structures common in educational assessments (Wainer & Kiely, 1987; Wainer & Lewis, 1990). Testlets are pervasive in language testing (reading passages, listening recordings, and integrated tasks) because they enhance authenticity and sampling of communicative contexts. Yet, testlets can also induce shared variance among item responses that is not fully attributable to the target ability (Bradlow et al., 1999; Wainer et al., 2007).

In language tests, testlet effects are often interpreted as reflecting a mixture of (a) stimulus-specific comprehension difficulty, (b) shared topical knowledge, (c) discourse-level coherence effects, and (d) strategy or memory effects that spread across items within the same stimulus (Buck, 2001; Goh & Vandergrift, 2022). From a measurement standpoint, these influences can be considered method variance or facet variance that needs to be modeled if it meaningfully changes score interpretation or precision (AERA et al., 2014; Kane, 2006).

The Rasch model is valued in language testing for its focus on fundamental measurement and invariant comparisons when model assumptions hold (Bond & Fox, 2015; Wright & Stone, 1979). Under the Rasch framework, items are assumed to measure a single trait with equal discrimination, and responses are expected to be locally independent. When LID is present, two broad problems arise: (1) parameter estimates may be biased, and (2) the test may appear more reliable than it truly is, because dependent items contribute redundant information (Hambleton et al., 1991; Marais & Andrich, 2008; Yen, 1984). Work on Rasch fit has long emphasized that acceptable mean-square indices do not guarantee that all assumptions hold, motivating complementary diagnostics and alternative model specifications (Smith, 2004; Wright et al, 1994; Yen, 1984).

To address dependence induced by clustered stimuli, researchers have developed testlet-aware extensions of IRT and Rasch models. Early work in IRT demonstrated that local dependence affects estimation and inference and proposed modeling strategies that treat clusters as meaningful residual dimensions (Thissen et al., 1989; Yen, 1993). Within the Rasch framework, the Rasch Testlet Model extends the unidimensional Rasch model by adding testlet-specific random effects (Wang & Wilson, 2005). This approach partitions variance into a general trait component (e.g., overall listening ability) and testlet variance capturing residual covariance due to shared stimuli, allowing improved parameter recovery and more realistic standard errors when dependence is present (DeMars, 2006; Wainer et al., 2007; Wang & Wilson, 2005).

Parallel developments in broader measurement include Bayesian random effects models for testlets (Bradlow et al., 1999), multidimensional comparisons that clarify formal relations among testlet models and alternative multidimensional specifications (Rijmen, 2009), and graphical loglinear model (Nielsen et al., 2025; Nielsen & O'Neill, 2025). In practice, these models frequently demonstrate improved global fit (deviance, AIC/BIC/CAIC reductions) when the assessment design plausibly induces dependence (DeMars, 2006; Rijmen, 2009; Wainer & Wang, 2001).

A central consequence of ignoring LID is inflated reliability: dependent items provide overlapping information, causing internal consistency and IRT-based precision indices to overstate how accurately a test measures the trait (Haberman, 2008; Marais & Andrich, 2008; Sireci et al., 1991). When testlet variance is modeled, reliability estimates often decline because redundant variance is reattributed from "true score" variance to method/testlet variance (DeMars, 2006; Wang & Wilson, 2005). Importantly, such reductions are typically interpreted as corrections that strengthen validity by aligning reported precision with the intended construct rather than design artifacts (AERA et al., 2014; Kane, 2006).

Much of the testlet literature in language assessment has focused on reading comprehension, where the persistent availability of text can facilitate cross-item strategy use and stimulate stronger within-passage dependencies (Aryadoust & Zhang, 2016; DeMars, 2006, Zolfaghari et al., in press). In contrast, listening differs in ways that may moderate the magnitude of dependence: input is transient, replay is typically not permitted in standardized settings, and memory demands constrain prolonged cross-item comparisons (Buck, 2001; Field, 2008; Goh & Vandergrift, 2022). These modality differences motivate an empirical question: does testlet dependence in listening exist at levels that materially affect model fit and score precision?

Comparative evidence from large-scale assessments suggests that listening sections often show lower inter-item correlations than reading sections, although dependence remains possible depending on item density, passage complexity, and item sequencing (Baghaei & Aryadoust, 2014; Lim, 2024; Weir & O'Sullivan, 2011). The implication is not that listening is free of LID, but that dependence may be smaller in magnitude and more contingent on design factors such as information load and question clustering.

Modern validity theory emphasizes that validation is an argument about score meaning and use, supported by evidence across content, internal structure, response processes, relations to other variables, and consequences (AERA et al., 2014; Kane, 2006, 2013; Mislevy, 1993). From this perspective, testlet modeling contributes primarily to internal-structure evidence and can also affect consequential validity insofar as it changes reported precision or fairness in high-stakes decisions.

Weir's (2005) evidence-based approach to language test validation similarly underscores the need to align test structure with construct theory and cognitive processing, supporting the view that psychometric modeling should be interpreted alongside theories of listening. Chapelle et al. (2008) further argue that validity arguments for large-scale language tests benefit from integrated evidence spanning psychometrics and cognition.

IELTS Academic Listening is organized into multiple recordings with sets of items per recording. While testlet effects have been widely studied in reading, relatively fewer studies have directly quantified and interpreted testlet variance in IELTS Listening using Rasch-based testlet modeling. Existing language testing research demonstrates that even modest testlet variance can improve fit under RTM and can meaningfully adjust reliability and standard errors (Baghaei & Aryadoust, 2014; Wainer et al., 2007; Wang & Wilson, 2005). Recent work continues to emphasize

regular evaluation of testlet effects as part of validation practice (Lim, 2024), motivating focused studies on listening modules specifically. The purpose of this study is to investigate whether a testlet effect exists in the IELTS Academic Listening test. Furthermore, the study compares the fit of two hypothesized models.

3. Methodology

3.1. Design and Data

The research employs a model-comparison approach to explore local item dependence in the IELTS Academic Listening module by comparing a unidimensional Rasch model to a Rasch testlet model. The test was an original official standard IELTS practice test (British Council, n.d.). The test included 40 items scored dichotomously, divided into four listening sections (one for each recording), mirroring the operational framework of IELTS Listening.

3.2. Participants and Setting

A total of 461 test takers participated in the IELTS Academic Listening Test, comprising 297 (64.4%) females and 164 (35.6%) males. They were selected from language learners of an adult language academy in Mashhad Iran. Their proficiency level was between B1 and B2 based on Common European Framework of Reference. Test takers needed to get IELTS score for immigration streams.

3.3. Procedure

Two measurement models were fitted: the first one was a Unidimensional Rasch Model (RM), which assumes local independence and equal item discrimination (Bond & Fox, 2015; Wright & Stone, 1979) and the second model was Rasch Testlet Model (RTM), which augments the RM by introducing testlet-specific random effects to account for within-testlet residual covariance (DeMars, 2006; Rijmen, 2009; Wainer et al., 2007; Wang & Wilson, 2005).

Both models estimate person ability on a logit scale and item difficulty parameters; the RTM additionally estimates variance components for each testlet (Bradlow et al., 1999; Wainer & Kiely, 1987; Wainer & Lewis, 1990).

Models were evaluated using marginal maximum likelihood with numerical integration for random effects (Adams et al, 2020; Bock & Aitkin, 1981; Rijmen, 2009). Global fit evaluations utilized deviance (-2LL) and information criteria (AIC, BIC, CAIC), where reduced values suggest superior parsimony-adjusted fit (Burnham & Anderson, 2002; Kass & Raftery, 1995). RM item fit was assessed using infit and outfit mean squares (Boone et al, 2014; Smith, 2004; Wright et al, 1994). Reliability was defined as the EAP/PV reliability for the overall dimension (Adams et al, 1997;

Mislevy, 1993; Warm, 1989). To assess LID magnitudes, we analyzed testlet variance components in RTM in relation to the variance of the general factor (DeMars, 2010; Haberman, 2008; Marais & Andrich, 2008; Wang & Wilson, 2005; Yen, 1984).

4. Results

4.1. Descriptive Statistics

The data of 461 test takers was collected and analyzed using IBM SPSS Statistics (Version 27). The listening test showed good internal consistency (Cronbach's $\alpha = .87$) according to Cronbach (1951). The test scores demonstrated a mean of $M = 21.56$ and a standard deviation of $SD = 1.26$. The mode of the scores was 5.5 and the score range extended from 2.0 to 8.5.

4.2. Item Fit under the Rasch Model

Models were estimated using the TAM (Test Analysis Moduls) package (Robitzsch et al., 2022) in the R statistical environment (R Core Team, 2022). All items met conventional infit/outfit expectations under the RM, indicating no gross misfit at the item level. Values between 0.5 and 1.5 considered acceptable (Linacre, 2022).

Table 1 presents item fit indices for all items, indicating that infit and outfit statistics are acceptable under RM.

Table 1

Rasch Model Item Difficulty Parameters and Fit Statistics for the Listening Items

Item	Difficulty	SE	Outfit	Infit
1	-1.56	.12	.90	.94
2	-.83	.10	1.02	1.04
3	.35	.10	1.05	1.04
4	-2.20	.14	1.16	1.07
5	-.91	.11	1.15	1.04
6	.49	.10	.92	0.96
7	-.42	.10	.98	0.98
8	.24	.10	1.18	1.11
9	1.97	.13	1.04	0.99
10	0.98	.11	.99	1.00
11	-2.35	.15	.80	0.92
12	-1.42	.11	.89	0.97
13	-2.19	.14	.84	0.95
14	-3.06	.19	1.07	0.99
15	-.22	.10	.79	0.83
16	-1.17	.12	.93	0.97
17	-1.29	.11	.85	0.93
18	-1.76	.13	.89	0.97
19	-3.07	.19	.91	0.96
20	-1.44	.12	.88	0.92
21	-.97	.11	.94	0.98
22	.52	.10	.92	0.94
23	-.17	.10	.93	0.94
24	-1.57	.12	.90	0.94
25	.12	.10	.81	0.85
26	-.24	.11	.89	0.92
27	.51	.10	1.30	1.23
28	-.72	.11	.92	0.96
29	.06	.11	.95	0.96
30	.85	.11	.98	1.00
31	-1.87	.13	1.05	1.04
32	.81	.11	1.01	1.00
33	1.78	.14	1.47	1.21
34	.25	.10	1.26	1.18
35	.08	.10	1.08	1.05
36	-.34	.10	.94	0.95
37	.56	.10	1.19	1.08
38	.11	.10	1.03	1.00
39	-1.19	.11	.84	0.93
40	-.41	.10	1.03	1.02

4.3. Testlet Variance Components (RTM)

As noted in Table2, RTM estimates showed non-zero testlet variances, with the general listening dimension variance near 0.989 and testlet variances of 0.292 (Testlet 1), 0.478 (Testlet 2), 0.050 (Testlet 3), and 0.133 (Testlet 4).

Table 2

Variances and Reliabilities of the General Listening Dimension and Testlet Dimensions

Dimension	Variance	Reliability
G	.989	.819
Testlet 1	.292	.255
Testlet 2	.478	.291
Testlet 3	.05	.056
Testlet 4	.133	.140

The variance of Testlet 2 is roughly half of the general dimension variance, and Testlet 1 is about one-third—both non-ignorable and indicative of moderate LID. The pattern is smaller than that typically observed in reading comprehension, plausibly due to the transient, acoustic nature of listening stimuli that limit prolonged cross-item interactions (cf. Buck, 2001; Field, 2008).

4.4. Model Fit Comparison: RM vs. RTM

Table 3 shows that the RTM achieved lower deviance (-2LL) and lower information criteria than the RM (e.g., -2LL: 18219.72 vs. 18287.55; AIC: 18310 vs. 18370; CAIC: 18541 vs. 18580; BIC: 18496 vs. 18539), indicating superior global fit when accounting for LID. This aligns with prior demonstrations that testlet or bifactor specifications typically improve parsimony-adjusted fit in testlet-based assessments (DeMars, 2006; Rijmen, 2009; Wainer et al., 2007; Wainer & Wang, 2001).

Table 3

Information Criteria and Reliability for the RM and RTM for the Listening Test

	-2LL	AIC	CAIC	BIC	Rel. G.	No. Para
Rasch Model	18287.55	18370	18580	18539	.857	41
Rasch Testlet Model	18219.72	18310	18541	18496	.819	45

Note: -2LL=deviance, Rel. G. =reliability of the general dimension, No. Para. = number of parameters

4.5. Reliability of the General Dimension

The reliability for the overall listening dimension fell from .857 (RM) to .819 (RTM) (Table 3). Such declines are anticipated when modeling residual dependence and indicate a reduction of artificially inflated precision under the local independence assumption (DeMars, 2006; Sireci et al., 1991; Wang & Wilson, 2005). Significantly, the decrease is slight, aligning with the moderate variances in testlets noted and the proposed smaller LID in listening compared to reading (Baghaei & Aryadoust, 2014; Bradlow et al., 1999).

Collectively, the findings suggest that the IELTS Listening module shows significant yet not excessively strong testlet effects. Modeling these effects with RTM enhances fit and results in a marginally reduced reliability for the general factor—an anticipated trade-off when distinguishing trait variance from method (testlet) variance. This fundamentally bolsters more justifiable score interpretations and aids in construct validity claims for listening evaluations (AERA et al., 2014; Kane, 2006; Weir & O'Sullivan, 2011).

5. Discussion

The current results verify that the IELTS Academic Listening section shows moderate testlet effects and that the Rasch Testlet Model offers a statistically better fit comparing to the unidimensional Rasch Model. This validates earlier studies claiming that neglecting local item dependence results in exaggerated reliability and skewed parameter estimates (DeMars, 2006; Sireci et al., 1991; Wainer & Wang, 2001; Yen, 1984). While each item aligns with the Rasch model, the significant variance components seen in the first two testlets indicate the persistence of shared stimulus effects, despite item-level fit indices seeming satisfactory—a phenomenon extensively documented in testlet literature (Bradlow et al., 1999; Rijmen, 2009; Wainer & Kiely, 1987).

In contrast to reading comprehension, the lesser degree of LID in listening is due to the fleeting, auditory aspect of input: audio is heard a single time and cannot be replayed, limiting chances for strategic item comparison (Buck, 2001; Field, 2008; Goh & Vandergrift, 2022). Comparable modality-specific variations have been noted in PISA, TOEFL, and Cambridge evaluations, with listening questions often demonstrating lower inter-item correlations compared to reading questions (Baghaei & Aryadoust, 2014; Lim, 2024; Weir & O'Sullivan, 2011). Consequently, the current findings build upon previous research by illustrating that, although LID is present in IELTS Listening, its effect is quantitatively less significant yet still psychometrically relevant.

Taking testlet variance into consideration, decreased overall reliability from .857 to .819 (see Table 3), follows a trend similar to reductions noted in other RTM uses (DeMars, 2006; Wang & Wilson, 2005). This reduction shows that the RM's reliability assessment was falsely elevated because of unaccounted dependencies (Haberman, 2008; Marais & Andrich, 2008; Zolfaghari et al., in press). The adjustment improves construct validity by making certain that reliability represents actual latent variance instead of method variance (AERA et al., 2014; Kane, 2006). In practical terms, the decrease of about .04 indicates moderate reliance, not strong enough to compromise total-score assessment but considerable enough to require recognition in validation records (Mislevy, 1993; Weir & O'Sullivan, 2011).

From a theoretical perspective, the research emphasizes that testlet design must be clearly represented when test items create contextual overlap. This corresponds with multidimensional measurement viewpoints where testlet factors indicate "method facets" in a bifactor structure

(Gibbons & Hedeker, 1992; Reckase, 2009; Reise, 2012). The RTM therefore connects conventional one-dimensional Rasch scaling with more adaptable hierarchical modeling, providing interpretive transparency while preserving Rasch's interval-scale benefits (Adams et al., 1997; Embretson & Reise, 2000).

From both pedagogical and operational perspective, the results highlight that characteristics of test design—including the quantity of items per recording, cognitive load, and information density—can influence the extent of LID (Buck, 2001; Sawaki et al., 2009). Minimizing the overcrowding of questions on one listening passage or varying item formats might alleviate reliance (Baghaei & Aryadoust, 2014; Boone et al., 2014). Furthermore, with the growth of automated delivery, adaptive listening assessments might include RTM-driven scoring to enhance accuracy across different testlet formats (Adams et al., 2020; Lim, 2024).

Aside from IELTS, the research supports the wider claim that the validation of language tests must incorporate both cognitive and psychometric data (Chapelle et al., 2008; Weir, 2005). Combining RTM results with cognitive listening models (Field, 2008; Goh & Vandergrift, 2022) reinforces a cohesive validity argument consistent with Kane's (2006) interpretive–use framework.

6. Conclusion

This study offers empirical proof that the IELTS Academic Listening module exhibits significant testlet effects, and that employing the Rasch Testlet Model to account for these effects results in improved statistical fit and more justifiable reliability estimates compared to a unidimensional Rasch model. The findings support prior conclusions that LID, even at low levels, can skew reliability and item parameters if not addressed (DeMars, 2006; Sireci et al., 1991; Wainer et al., 2007).

Moreover, these results support the regular assessment of LID in large listening evaluations and the inclusion of testlet modeling in both research and practical scoring. Although the overall psychometric integrity of the IELTS Listening test is strong, recognizing and addressing LID improves clarity and equity.

Future studies can investigate the issue with bigger and more varied samples, broaden modeling to additional language areas (e.g., combined listening–speaking activities), and investigate Bayesian or multidimensional RTM alternatives to understand intricate dependency trends (Bradlow et al., 1999; Lim, 2024; Rijmen, 2009). Ongoing advancements in this direction will enhance the validity argument for IELTS and similar listening evaluations, guaranteeing that the reported scores accurately reflect candidates' listening skills instead of being influenced by specific testlet methods (AERA et al., 2014; Kane, 2013).

References

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*(1), 1–23. https://doi.org/10.1177/0146621697211001

Adams, R. J., Wu, M. L., Cloney, D., Berezner, A., & Wilson, M. (2020). *ACER ConQuest: Generalised item response modelling software* (Version 5.29) [Computer software]. Australian Council for Educational Research. https://www.acer.org/au/conquest

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. AERA.

Aryadoust, V., & Zhang, L. (2016). Fitting the mixed Rasch model to a reading comprehension test: Exploring individual difference profiles in L2 reading. *Language Testing*, *33*(4), 529–553. https://doi.org/10.1177/0265532215594640

Baghaei, P., & Aryadoust, V. (2014). Modeling local item dependence due to common test format with a Multidimensional Rasch Model. *International Journal of Testing*, *15*(1), 71–87. https://doi.org/10.1080/15305058.2014.941108

Baghaei, P., Ravand, H., & Striethолt, R. (2025). Modeling Local Item Dependence in PIRLS 2016: Comparing ePIRLS and Paper-Based PIRLS using the Rasch Testlet Model. *Psychological Test and Assessment Modeling*, 67, 127–140. https://doi.org/10.2440/001-0021

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. https://doi.org/10.1007/BF02293801

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*, (3rd ed.). Routledge. https://doi.org/10.4324/9781315814698

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. Springer Netherlands. https://doi.org/10.1007/978-94-007-6857-4

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*(2), 153–168. https://doi.org/10.1007/BF02294533

British Council. (n.d.). *Free online IELTS Listening practice tests*. Take IELTS. https://takeielts.britishcouncil.org/take-ielts/prepare/free-ielts-english-practice-tests/listening

British Council. (n.d.). *IELTS test format explained.* https://www.britishcouncil.org/

Buck, G. (2001). *Assessing listening*. Cambridge: Cambridge University Press. doi: 101017/CBO9780511732959

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference* (2nd ed.). Springer. https://doi.org/10.1007/b97636

Cambridge English. (n.d.). *IELTS test format.* https://www.cambridgeenglish.org/

Chapelle, C.A., Enright, M.K., & Jamieson, J.M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language* ™ (1st ed.). Routledge. https://doi.org/10.4324/9780203937891

Cronbach, L. J. (1951). Coefficient Alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. https://doi.org/10.1007/BF02310555

DeMars, C.E. (2006), Application of the Bi-Factor Multidimensional Item Response Theory Model to Testlet-Based Tests. *Journal of Educational Measurement*, *43*(2), 145-168. https://doi.org/10.1111/j.1745-3984.2006.00010.x

DeMars, C. (2010). *Item response theory.* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195377033.001.0001

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists* . Psychology Press. https://doi.org/10.4324/9781410605269

Field, J. (2008). *Listening in the language classroom.* Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511575945

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bifactor analysis. *Psychometrika, 57*(3), 423–436. https://doi.org/10.1007/BF02295430

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

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics, 33*(2), 204–229. https://doi.org/10.3102/1076998607302636

Hagell, p. (2026). Local dependence in health outcome measurement: The case of the 8-item Parkinson 's disease questionnaire (PDQ-8). Educational Methods & Psychometrics, 4 (SAMC 2024 Special Issue): 27. https://doi.org/10.65301/emp.2026.256

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Sage Press.

IELTS. (n.d.). *IELTS Academic Listening format.* https://www.ielts.org/

Kane, M. T. (2006). Validation. In *R. B. Educational Measurement* (4th ed., pp. 17-64). American Council on Education/Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.2307/23353796

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Lim, Y.S. (2024). Assessment of Testlet Effects: Testing it all at once. In: Hwang, H., Wu, H., Sweet, T. (eds) *Quantitative Psychology. IMPS 2023. Springer Proceedings in Mathematics & Statistics*, 452. Springer, Cham. https://doi.org/10.1007/978-3-031-55548-0_26

Linacre, J. M. (2022). *A user's guide to Winsteps Rasch-model computer programs* (v.5.2.3). Winsteps.com.

Marais, I., & Andrich, D. (2008). "Effects of varying magnitude and patterns of response dependence in the unidimensional Rasch model." *Journal of Applied Measurement*, *9*(2): 105-124.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational Measurement (3rd ed.), (pp. 13-103). New York: Macmillan Publishing Co.

Nielsen, T., & O'Neill, L. D. (2025). Critical thinking at entry to university: Rasch modeling of a generalized version of the critical thinking scale of the Motivated Strategies for Learning Questionnaire in a multi-program context. *Educational Methods & Psychometrics*, *3*, Article 21. https://doi.org/10.61186/emp.2025.8.

Nielsen, T., Fellinghauer, C., Strobl, C., Kronthaler, D., & Kreiner, S. (2025). Statistical anxiety and attitudes towards statistics in psychology students: A measurement comparison study of the German and Danish language versions of the HFS-R. *Educational Methods & Psychometrics*, *3*, Article 19. https://doi.org/10.61186/emp.2025.6

R Core Team. (2022). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.r-project.org/

Reckase, M. D. (2009). *Multidimensional item response theory*. Springer. https://doi.org/10.1007/978-0-387-89976-3

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*, 667-696. https://doi.org/10.1080/00273171.2012.715555

Rijmen, F. (2009). Three multidimensional models for testlet-based tests: Formal relations and an empirical comparison. *British Journal of Mathematical and Statistical Psychology, 62*(1), 13–36. https://doi.org/10.1002/j.2333-8504.2009.tb02194.x

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2022). *TAM: Test analysis modules* [R package, version 4.1-1]. https://CRAN.R-project.org/package=TAM

Sawaki, Y., Stricker, L., & Oranje, A. (2009). Factor structure of the TOEFL iBT reading section. *Language Testing, 26*(1), 5–30. https://doi.org/10.1177/0265532208097335

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*(3), 237–247. https://doi.org/ 10.1111/j.1745-3984.1991.tb00356.x

Smith, R. M. (2004). Detecting item bias with the Rasch Model. In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 301–317). JAM Press.

Thissen, D., Steinberg, L. & Mooney, J.A. (1989), Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement*, *26*: 247-260. https://doi.org/10.1111/j.1745-3984.1989.tb00331.x

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications.* Cambridge University Press.

Wainer, H. and Kiely, G.L. (1987), Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*, 185-201. https://doi.org/10.1111/j.1745-3984.1987.tb00274.x

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement, 27*(1), 1–14. https://doi.org/10.1111/j.1745-3984.1990.tb00730.x

Wainer, H., & Wang, X. (2001). *Using a new statistical model for testlets to score TOEFL*. ETS. Princeton

Wang, W.-C., & Wilson, M. (2005). The Rasch Testlet Model. *Applied Psychological Measurement, 29*(2), 126–149. https://doi.org/10.1177/0146621604271053

Warm, T. A. (1989). Weighted likelihood estimation of ability. *Psychometrika, 54*, 427–450. https://doi.org/10.1007/BF02294627

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan. https://doi.org/10.1057/9780230514577

Weir, C. J., & O'Sullivan, B. (2011). *Test development and validation*. Palgrave Macmillan.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. MESA Press.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*(2), 125–145. https://doi.org/10.1177/014662168400800201

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187–213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x

Zhu, X., Raquel, M., & Aryadoust, V. (2019). Structural equation modeling to predict performance in English proficiency tests. In V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment: Volume II. Advanced methods* (pp. 101–126). Routledge. https://doi.org/10.4324/9781315187808-5

Zolfaghari, N. R., Baghaei, P., Seifoori, Z. (in press). Examining the dimensionality of IELTS academic reading: A Rasch testlet model approach to local item dependence. *International Journal of Language Testing*. https://www.ijlt.ir/