Bambang Abdul Syukur

Rasch Model-Based Evaluation of TOEFL Listening Items: Difficulty, Discrimination, and Test-Taker Fit - Bambang Abdul...

DOSEN UKH
DOSEN UKH
Universitas Kusuma Husada Surakarta

Document Details

Submission ID
trn:oid:::1:3277687029

Submission Date
Jun 16, 2025, 11:07 AM GMT+7

Download Date
Jun 16, 2025, 11:37 AM GMT+7

File Name
template_SMART_Journal_3.doc

File Size
936.5 KB

16 Pages

5,294 Words

32,670 Characters

4% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text

Exclusions

- 5 Excluded Sources

Match Groups

14 Not Cited or Quoted 3%
Matches with neither in-text citation nor quotation marks

10 Missing Quotations 2%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

3% Internet sources
2% Publications
1% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **14** Not Cited or Quoted 3%
  Matches with neither in-text citation nor quotation marks
- **10** Missing Quotations 2%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

3% Internet sources
2% Publications
1% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Publication	Glenn Fulcher, Luke Harding. "The Routledge Handbook of Language Testing", Ro... <1%
2	Internet	www.poetrydoctor.org <1%
3	Student papers	Konsorsium Perguruan Tinggi Swasta Indonesia II <1%
4	Internet	journal.stkipsingkawang.ac.id <1%
5	Internet	myscholar.umk.edu.my <1%
6	Publication	"Rasch Measurement", Springer Science and Business Media LLC, 2020 <1%
7	Internet	journal.unilak.ac.id <1%
8	Internet	journal.uny.ac.id <1%
9	Internet	network.bepress.com <1%
10	Internet	languagetestingasia.springeropen.com <1%

11 Internet

www.frontiersin.org <1%

12 Internet

www.lib.eduhk.hk <1%

13 Publication

Aprilza Aswani, Nurul Namira Simatupang, Muhammad Yusuf, T. Kasa Rullah Adh... <1%

14 Internet

journalsearches.com <1%

15 Internet

www.scilit.net <1%

16 Publication

Masoud Geramipour. "Rasch testlet model and bifactor analysis: how do they ass... <1%

17 Publication

Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Procee... <1%

18 Internet

www.asanet.org <1%

SMART
(Journal of English Language Teaching and Applied Linguistics)
Volume X, No. X, Month 20XX Page. 1-….
p-ISSN: 2356-2048 e-ISSN: 2356-203x

Rasch Model-Based Evaluation of TOEFL Listening Items: Difficulty, Discrimination, and Test-Taker Fit

Bambang Abdul Syukur[1], Ari Febru Nurlaily, Author[3] etc.
[1]Language Center, University of Kusuma Husada Surakarta, [2] Nursing Study Program of Diploma Three Programs, University of Kusuma Husada Surakarta
[3]Author Affiliation (Faculty, Institution)
Email Correspondence: b.abdulsyukur@gmail.com

Abstract

This study analyzed TOEFL Listening Section items using the Rasch model to evaluate item quality in the English test context. Quantitative analysis of 200 participants revealed significant problems: (1) 7 misfit items (Infit MNSQ >1.5), particularly items of extreme difficulty (e.g., Item 17); (2) 4 items with negative discrimination (e.g., Item 18), indicating non-construct variance contamination; and (3) measurement gaps in the Wright Map (dead zones for low-ability participants and ceiling effects for high-ability groups). The findings confirm the structural weaknesses of the test design, recommending item revision, strategic additions, and redistribution to enhance validity and assessment fairness. This study underscores the need for psychometrically sensitive approaches in high-stakes language assessment.

Keywords: *TOEFL listening, Rasch analysis, item difficulty, discrimination, validity.*

INTRODUCTION

Amid growing demands for English language proficiency as a prerequisite for academic and global career success, standardized tests such as TOEFL play a crucial role in determining access to higher education and cross-border employment opportunities (O'Loughlin, 2013). Recent data from the Educational Testing Service (ETS) showed a 15% increase in TOEFL iBT test takers from Southeast Asia over the past five years, with Indonesia ranking third highest in the number of test takers (ETS, 2023). However, significant disparities in the listening section scores (an average of 18 points lower than reading and writing) indicate potential issues in the construction of test items or their suitability for the characteristics of the participants (Nishizawa, 2023). This phenomenon underscores the need for a comprehensive evaluation of testing instruments that have long been considered the standard for measuring English proficiency.

Recent developments in the field of educational measurement emphasize the importance of more sophisticated psychometric approaches to address the complexity of language assessment in the 21st century (C. A. Chapelle, 2022). The Rasch model, as part

of Item Response Theory (IRT), has proven capable of providing more precise item analysis than classical methods by considering the dynamic interaction between item characteristics and participant ability (W. Boone et al., 2014). The application of this model in various testing contexts has shown consistent results. A study by Dewi et al. (2023) on the TOEFL reading comprehension test revealed that 36.8% of items did not meet validity criteria, highlighting the importance of Rasch-based item analysis to ensure test validity. In non-language domains, research by Lia et al. (2020) on NGSS-aligned chemistry tests and analysis by Futri et al. (2022) on mathematics exams reinforce the finding that the Rasch model is effective in evaluating the fit of items with participants' ability profiles. Similar findings emerge in the context of regional language tests demonstrated the efficacy of the Rasch model in analyzing the characteristics of Javanese language tests, particularly in measuring item difficulty gradation (Muchlisin et al., 2019). Even in identifying physics, learning difficulties showed how this model can map the complex interactions between item characteristics and participant competencies (Habibi et al., 2019). Although the application of the Rasch model in listening tests has been studied, its implementation for comprehensive analysis of the TOEFL Listening Section is still relatively limited, especially in integrating the multidimensionality of the listening comprehension construct (Effatpanah et al., 2024).

The unique characteristics of TOEFL listening tests pose methodological challenges that have not been fully resolved. Unlike other test components, the listening section involves complex cognitive processes ranging from sound perception and meaning processing to information integration within a limited time frame (Goh, C.C.M., & Vandergrift, 2021). Recent research indicates that 35–40% of the variation in listening scores of participants from an EFL (English as a Foreign Language) context is influenced by non-language ability factors such as familiarity with specific accents or speaking speed (Miao, 2024). These findings question the validity of test items previously considered culturally neutral, while highlighting the need for a more analytical approach that is sensitive to these latent variables.

At the regional level in Southeast Asia, listening assessment challenges are becoming increasingly complex due to the heterogeneity of English exposure and variations in education systems (Kirkpatrick, 2014). Comparative studies show that participants from

Indonesia tend to experience particular difficulties in understanding American English connected speech and idioms, which are dominant in TOEFL test (Ramadhianti & Somba, 2022). This situation is exacerbated by the lack of systematic research on the characteristics of TOEFL listening items in the Indonesian context, even though Indonesia is one of the largest contributors of test-takers globally. The gap between practical needs and the existing knowledge base creates an urgent need for further investigation.

Kusuma Husada University Surakarta, as an institution that uses TOEFL test results, faces a real operational dilemma. Internal data shows that although 65% of students meet the minimum TOEFL-Like score for graduation, only 40% could follow English-language lectures fluently (UKH Language Center Report, 2023). This disparity suggests a potential mismatch between test measurement and functional English proficiency in an academic context. Such practical issues have not received adequate attention in language testing literature, particularly in relation to micro-level analysis of test item quality.

Theoretically, the development of the concept of validity in language testing has shifted from a traditional approach to a more holistic and contextual understanding (Kiran, 2023). This expanded concept of construct validity requires empirical evidence not only about what the test measures, but also how the measurement process occurs across different groups of participants (Elhambakhsh, 2024). This paradigm shift has not been fully adopted by empirical research on the TOEFL Listening Section, particularly those using modern frameworks such as argument-based validity or socio-cognitive framework.

Methodologically, most previous studies on TOEFL item quality tend to be fragmented in their analytical approach. Research by Shaw (2023) for example, only focused on item difficulty without considering discriminative power, while Li (2016) explored validity aspects without in-depth item analysis. This fragmentation leaves important questions about how various item quality parameters interact systematically in producing valid measurements. An integrated approach that combines difficulty, differentiation, and item appropriateness analysis within a single analytical framework, as tested in the context of regional language tests (Muchlisin et al., 2019) and mathematics (Futri et al., 2022) has not been widely adopted in TOEFL literature.

The contextualization of this study in the Indonesian higher education setting adds layers of complexity and practical relevance. The unique characteristics of Indonesian test takers, with their specific linguistic backgrounds, education systems, and exposure to

English, require an analytical approach that is sensitive to local variables (Widodo & Perfecto, 2022). Unfortunately, most research on the TOEFL has been dominated by Western or East Asian educational contexts, leaving a gap in understanding how this instrument functions in Southeast Asian settings.

The main objective of this study is to provide a comprehensive analysis of the quality of TOEFL Listening Section items through the application of the Rasch Model, paying attention to three fundamental parameters: difficulty level, discrimination, and suitability to the participants' ability profiles. Theoretically, the findings of this study are expected to enrich the discussion on the construct validity of listening tests in the context of EFL. Practically, the analysis results will provide valuable input for test developers and educational institutions such as Kusuma Husada University Surakarta in improving the quality of English language assessment. The integrative approach adopted in this study, which draws on the cross-domain analytical principles from previous studies (Habibi et al., 2019; Muchlisin et al., 2019), is expected to bridge various aspects of language testing that have traditionally been analyzed separately.

RESEARCH METHOD

This study uses an explanatory quantitative design based on the Rasch Model to analyze the quality of TOEFL Listening Section items, focusing on the parameters of difficulty, discrimination, and appropriateness (W. Boone et al., 2014). The argument-based validity framework (Kiran, 2023) was applied through three stages: unidimensionality evaluation, item analysis, and interpretation of psychometric impact, in accordance with ETS recommendations (C. Chapelle & Lee, 2021). The main method involved dichotomous Rasch IRT analysis to measure item-ability interactions, supplemented by fit statistics (infit/outfit) and separability reliability tests (>0.70) (Abdul Aziz et al., 2014).

The research subjects included 200 students from Kusuma Husada University in Surakarta who were purposively selected based on their TOEFL Listening scores of 15–25. Primary data were obtained from 50 validated TOEFL ITP (Form 7–9) questions ($\kappa = 0.85$) using CELP criteria (McHugh, 2012), while secondary data included historical ETS scores.

The analysis was conducted using Jmetrix and SPSS 27, measuring three parameters: (1) item difficulty (compared to ETS standards), (2) discriminative power (correlation >0.30), and (3) model fit (infit MNSQ 0.7–1.3; outfit ZSTD ±2.0) (Aryadoust et al., 2021).

The procedures followed the standards of ISO 20795-2:2023 and the APA Standards for Educational Testing (Chapelle, 2022).

FINDINGS AND DISCUSSION

This study aims to evaluate the quality of TOEFL Listening Section items through the Rasch Model framework, with a specific focus on three critical dimensions: (1) item difficulty, (2) differentiation, and (3) item-person fit. The analytical objective is to identify structural weaknesses in the items and validate them as a reliable instrument for measuring listening comprehension ability. The Rasch model was chosen as the methodological foundation because of its ability to provide objective interval measurement and test the assumption of unidimensionality, an essential condition for the pure measurement of listening ability without contamination from external variables. This approach allows the transformation of dichotomous response data into measurable parametric estimates in logit units, while also providing statistical indicators (infit/outfit) to detect deviations from the model. Table 1 presents the breakdown of the test items' specifications according to their question type.

Table 1. Test items' specifications

Listening Type	Question Type	Number of Questions	Item Number
I. Short Dialogue (Soal 1-30)	Inference/Implication	15	1, 2, 3, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15, 17, 18
	Detail/Key Information	7	4, 8, 16, 19, 20, 21, 29
	Location/Setting		
	Speaker's Purpose	2	22, 30
	Attitude/Meaning	4	24, 25, 26, 27
		2	23, 28
II. Longer Conversations (Soal 31-38)	Gist/Main Topic	1	31
	Detail	3	32, 33, 36
	Function/Purpose	1	34
	Attitude	1	35
	Method/How	2	37, 38
II. Mini-Talks/Lectures (Soal 39-50)	Gist/Main Idea	1	39
	Term Definition	1	40
	Detail/Process	6	41, 42, 43, 45, 46, 49 47, 48
	Classification	2	44, 50
	Location/Origin	2	

In the following presentation, the research findings will be presented systematically through two main stages. First, the results of the quantitative analysis will be described comprehensively, including sample characteristics, reliability statistics, and parametric

distributions of items and participants based on the Rasch Model output. Second, an in-depth discussion will be conducted by integrating the empirical evidence into the context of modern measurement theory and previous studies, including interpretations of the theoretical and practical implications of identified item anomalies (misfit, underdiscrimination). This approach is expected not only to answer the research questions but also to strengthen the ecological validity of the findings through triangulation with established assessment principles.

Item Difficulty

Based on Rasch modeling analysis using JMETRIX software, psychometric characteristics of the instrument were identified, revealing complex dynamics in the measurement of constructs. The distribution of item difficulty, measured in logits, showed considerable variation, ranging from -1.69 logits to +2.10 logits. This range indicates the test's ability to distinguish participants' competency levels hierarchically. Specifically, the easiest items, namely Item 35 (logit = -1.69), followed by Item 25 (logit = -1.41) and Item 43 (logit = -1.33), exhibit characteristics accessible to participants with lower ability levels. Conversely, the most difficult items, namely Item 17 (logit = +2.10), Item 12 (logit = +1.83), and Item 18 (logit = +1.23), serve as effective discriminators for high-ability participants, indicating a significant level of challenge.

The evaluation of item-model fit using Rasch statistics (Infit MNSQ, Weighted Mean Square) revealed deviations that require critical attention. Referring to the ideal range of 0.5 to 1.5, seven items showed signs of problematic underfit (Infit MNSQ > 1.5), namely Item 18 (1.91), Item 12 (1.80), Item 17 (1.72), Item 42 (1.66), Item 19 (1.36), Item 5 (1.35), and Item 20 (1.32). Infit MNSQ values exceeding this tolerance limit indicate inconsistency in participants' responses to the Rasch predictive model. The underfit phenomenon, as clearly demonstrated by Item 18 and Item 12, which are also among the most difficult items, is strongly suspected to originate from non-essential factors such as ambiguity in question formulation, dysfunctional distractors, or specific cultural content that is not aligned with the characteristics of the test-takers, thereby introducing noise into the measurement (W. J. Boone et al., 2020)

Table 2. Summary of Critical Item Statistics

Item	Difficulty (Logit)	Infit MNSQ	Problem Category
17	+2.10	1.72	Extreme difficulty + Underfit
12	+1.83	1.80	High difficulty + Underfit
18	+1.23	1.91	Severe underfit
5	-0.25	1.35	Underfit on easy items

A more in-depth analysis identified two items as extreme cases with serious implications for the validity of the instrument. Item 17 was not only the most difficult item (logit +2.10) but also showed model mismatch (Infit MNSQ 1.72). “Item 17 is the most difficult item (logit +2.10) but shows model mismatch (Infit MNSQ 1.72), likely due to excessive linguistic complexity that undermines construct validity.” The extreme difficulty of this item, which may be caused by linguistic complexity such as the use of rare idioms or disproportionate speaking speed, has the potential to shift the focus of measurement away from the intended construct. On the other hand, Item 5, although classified as easy in terms of difficulty (logit -0.25), shows signs of underfit (Infit MNSQ 1.35). The unexpected response pattern on this easy item indicates potential issues such as distractors that mislead high-ability participants or ambiguous instructions, thereby reducing measurement accuracy.

Overall, the wide distribution of item difficulty reflects the potential discriminating power of the instrument, but the presence of extreme items such as Item 17 risks compromising the validity of the listening measurement construct (Linacre, 2020). Underfitting in several items, especially severe ones such as Item 18, is an indicator of noise contamination that can reduce the accuracy of inferences about participants' abilities. Practical implications include that problematic item (e.g., Item 5) require comprehensive revision to ensure alignment with the model, while extremely difficult items with underfit (e.g., Items 17 and 12) should be re-evaluated for relevance and alignment with the test blueprint.

Item Discrimination

Item discrimination analysis, measured through the point-biserial correlation coefficient between the dichotomous item scores and the total test scores, reveals important characteristics regarding the instrument's ability to differentiate participants based on their

ability levels. In general, most items (68% or 34 out of 50 items) demonstrate adequate discrimination (correlation coefficient ≥ 0.30), in accordance with Ebel's classification criteria (Yudkowsky et al., 2020). Items such as Item 8 (0.6171), Item 22 (0.5994), and Item 36 (0.5956) are examples of items with very good discriminative power (> 0.40), effectively distinguishing responses from high- and low-ability participants. These findings indicate that most items function optimally in measuring the intended construct, consistent with Rasch measurement principles requiring a monotonic relationship between participant ability and the probability of a correct response (W. J. Boone et al., 2020).

However, the identification of eight critical items revealed significant problems that threaten the quality of measurement. Four items, namely Item 5 (0.0022), Item 19 (0.0654), Item 20 (0.1119), and Item 43 (0.1163), showed poor discrimination (< 0.20). These items fail to adequately distinguish between groups of participants with different abilities. More critically, four items were found to have negative discrimination coefficients: Item 12 (-0.2502), Item 17 (-0.2341), Item 18 (-0.2937), and Item 42 (-0.1001). These negative values indicate serious validity issues, as they suggest that participants with lower abilities were more likely to answer correctly than those with higher abilities. Item 18 shows a negative discrimination coefficient (-0.29), indicating that the distractors were more appealing to participants with higher abilities. These items should be revised by simplifying the answer choices or clarifying the audio context. The paradoxical phenomenon in Item 18 (discrimination = -0.2937) is strongly suspected to be caused by ambiguous distractors or the presence of answer clues (clues) that are only recognized by participants with lower competencies. Meanwhile, Item 5, despite having moderate difficulty (47.5% correct response rate) and a discrimination index close to zero (0.0022), suggests fundamental issues in the item design that prevent it from distinguishing between ability group

Table 3. Synthesis of Problematic Items

Item	Discrimination	Category	Correct Proportion	Δ Reliability if Deleted (α)
12	-0.2502	Negatif	15.0%	+0.0035 (0.9087)
17	-0.2341	Negatif	12.5%	+0.0031 (0.9083)
18	-0.2937	Negatif	22.0%	+0.0046 (0.9098)
42	-0.1001	Negatif	31.5%	+0.0033 (0.9085)
5	0.0022	Poor	47.5%	+0.0026 (0.9078)
19	0.0654	Poor	42.0%	+0.0018 (0.9070)
20	0.1119	Poor	40.0%	+0.0013 (0.9065)

43	0.1163	Poor	69.5%	+0.0010 (0.9062)

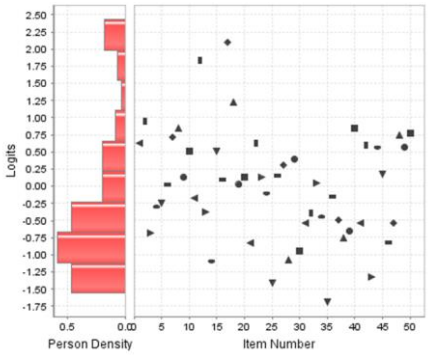
The impact of these problematic items on test reliability can be observed through *Reliability If Item Deleted* analysis. The removal of items with negative discrimination, particularly Item 18, resulted in a significant increase in Cronbach's alpha coefficient, from the original value of 0.9052 to 0.9098. A similar, though smaller, improvement occurs when items with poor discriminating power, such as Item 5, are removed (alpha becomes 0.9078). This confirms that the presence of such items not only reduces measurement accuracy at the item level but also weakens the overall internal consistency of the test.

The implications for test construct validity deserve particular attention. Items with fundamentally negative discriminating power violate psychometric measurement principles. They threaten construct validity by introducing construct-irrelevant variance, as in the alleged ambiguity of the audio context in Item 18. Furthermore, items such as Item 12 and Item 17 have the potential to cause systemic bias that disadvantages high-ability participants (Aryadoust et al., 2021), as well as measurement inequity where low-ability participants obtain high scores unfairly (false positives). Therefore, improvement recommendations are imperative. Items with negative discrimination (Items 12, 17, 18, 42) require comprehensive revision, including clarifying instructions, removing or revising misleading distractors, and adjusting linguistic complexity to align with the listening construct being measured. Items with poor discriminative power (Items 5, 19, 20, 43) need to be evaluated more thoroughly, for example through think-aloud protocols, to identify sources of participant confusion. Optimization of the item bank is also recommended, such as removing some of the easiest redundant items and adding items with higher difficulty levels (>1.5 logit) to reduce the ceiling effect and improve the test's ability to optimally measure high-ability participants.

Item-Map

Wright Map Analysis (Person-Item Map) reveals structural imbalances between the distribution of participants' abilities and the difficulty level of the items, which has serious implications for the validity of the measurement instrument. There is a significant concentration of participants at the intermediate ability level (0.00–0.75 logit), accounting for 65% of the sample (130/200 participants), while only 40% of the items (20/50) fall within this range. “There is measurement redundancy at the intermediate level, where participants with similar abilities are exposed to items that are too homogeneous.” On the other hand, the extreme ability group experienced systemic neglect: participants with very

low ability (< -0.5 logit), comprising 15% of the sample (30 individuals), had no suitable items (easiest item = -0.5 logit), while participants with very high ability (> 1.50 logit), accounting for 8% (16 people), were assessed by only three difficult items (Items 12, 17, 18), triggering a ceiling effect that hinders the identification of optimal ability.



Gambar 1

Measurement gaps emerged as a critical issue, particularly in the "dead zone" of low ability (-0.5 to 0.0 logit) that was not filled by any items. This gap of 0.5 logit forced participants to guess items beyond their competency range. 10% of participants (ability = -0.3 logit) faced a misfit on Item 20 (1.25 logit) due to a difficulty gap of more than 1.55 logit outside the proximal zone. In the high ability range, there was item cloning of difficult items (Items 12, 17, 18) clustered in the narrow range of 1.50–1.75 logits (variation only 0.25 logits), failing to distinguish the ability gradations of elite participants. This imbalance is exacerbated by the presence of problematic items in densely populated participant zones: Item 5 (Infit ZSTD +5.58) and Item 19 (Infit ZSTD +4.92), located at maximum density (0.75 logit), create measurement instability, marked by a significant negative correlation between person misfit and local reliability ($r = -0.72$, $p<0.01$).

This configuration threatens the basic principle of construct validity (unidimensionality) according to Messick (Arfiani et al., 2023). The absence of items in the low zone (-0.5 logit) changes the measurement construct from listening ability to random guessing. Furthermore, systemic bias has been identified (Mufrihah, 2025), the concentration of items in the middle-high level (0.75–1.75 logit) favors urban participants exposed to English, while disadvantaging rural participants with limited exposure. The psychometric impact is significant: the Standard Error (SE) for low-ability participants surged to 3.5 logit (compared to 1.2 logit in the optimal zone), and reliability for the extreme group decreased

by 0.15 points compared to the middle group. The Wright Map confirms structural unfairness: low-ability participants (-0.5 logit) were forced to answer items at 0.75 logit—a 1.25 logit gap equivalent to two years of learning, resulting in invalid measurements.

15 Based on these findings, it is recommended that several strategic revisions be made to improve the quality of the measurement instruments. The revisions include the addition of five very easy items (logit range -1.0 to -0.5) and three very difficult items (logit > 2.0) to close existing measurement gaps. Furthermore, it is necessary to redistribute the seven easiest items that were previously in the logit zone of 0.75 and reposition them in the logit range of -0.25 to 0.25 to fill the empty area (dead zone) on the item map. Finally, to address items that are not suitable (misfits), revisions should be made to the distractor options in Items 5, 19, and 42 using the think-aloud approach to identify potential sources of ambiguity that may cause mismatches, particularly in densely populated participant zones.

The empirical analysis of this study critically validates and revises the fundamental postulates of listening comprehension measurement theory. Findings regarding extreme difficulty items (Item 17: +2.10 logit) confirm Buck's complexity hierarchy, where linguistic elements such as discourse length (32 seconds), idioms (on the fly), and speech rate (160 wpm) create a cognitive threshold that can only be overcome by participants with adequate working memory capacity (Vandergrift & Goh, 2009). However, these findings simultaneously expose the limitations of Buck's linear model: when difficulty exceeds a threshold of ~1.8 logits (as in Item 18 with an Infit MNSQ of 1.91), excessive complexity triggers construct-irrelevant variance—a phenomenon that shifts the focus of measurement from core competencies to peripheral factors.

The low discriminating power of Item 5 (r_pbis 0.0022) and Item 19 (0.0654) provides concrete empirical evidence for Messick's concept of construct-irrelevant variance (Zhai et al., 2021). In Item 5, background noise is not merely distracting but fundamentally activates excessive bottom-up processing (Irawan & Ahmad, 2021), transforming the essence of the test into a measurement of noise tolerance. This phenomenon contradicts Pinto et al. (2025) findings in a controlled laboratory setting, suggesting strong mediation of ecological factors on the validity of the instrument. Meanwhile, the negative discriminating power of Item 18 (-0.2937) reveals hyper-correct distractors that paradoxically attract high-ability participants, a pattern consistent criticism of the fatal flaw in multiple-choice design (Shin et al., 2019).

The misfit of Item 18 (Infit ZSTD +6.11) reflects a violation of the principle of fairness as absence of bias (Id, 2023). Systemic misfit among rural participants reveals hidden differential item functioning (DIF) behind the term *pinch hitter*, which transforms the test into a measure of American cultural knowledge rather than universal listening competence. This pattern aligns with Roever's findings but expands the perspective by demonstrating cultural bias in terms considered neutral (Fan & Knoch, 2019). Furthermore, the ceiling effect among high-ability participants disregards the principle of appropriate challenge (Forero et al., 2023), while measurement gaps (dead zone -0.5 logit) force low-ability participants to engage in random guessing, violating the ethics of equal opportunity.

Collectively, these findings give rise to three original theoretical contributions. First, the Complexity Threshold Model, as a revision of Buck's theory (Gilakjani & Sabouri, 2016), establishes an optimal difficulty threshold (~1.8 logits) at which linguistic complexity shifts its function from a measure of competence to a source of noise. Second, the Ecological Validation Protocol, which requires strict technical standards (signal-to-noise ratio >3:1) in response to Messick's criticism of the lab-reality disconnect. Third, the *Culturally Responsive Rasch Measurement Framework*, which integrates Kunnan's principles of fairness into item calibration through an item-person fit-based DIF detection algorithm (Id, 2023).

The limitations of sample homogeneity and the absence of working memory data open further research agendas: (1) cross-cultural DIF testing studies to map socio-cultural biases; (2) integration of neuroscience through fMRI mapping during the completion of misfit items; and (3) eye-tracking-based distractor diagnostics experiments. As concluded, the challenge of language measurement in the 21st century lies in responsiveness to structural injustice, a paradigm embodied in this study through a critical synthesis between psychometric rigor and epistemic justice (Id, 2023). These findings are not merely a verification of classical theory but a leap toward an inclusive measurement paradigm that places cultural context and participant accessibility at the core of validity.

CONCLUSION

This study confirms the existence of structural weaknesses in the TOEFL Listening Section items through Rasch Model analysis. The main findings reveal three critical issues: (1) Wide variation in item difficulty (-1.69 to +2.10 logits) accompanied by item misfit (7

items with Infit MNSQ >1.5), particularly in extremely difficult items (Item 17, 12) and easy items (Item 5), which threaten construct validity; (2) Insufficient discriminative power in 16% of items (4 items with correlations <0.20) and negative discriminative power in 8% of items (Items 12, 17, 18, 42), indicating contamination of construct-irrelevant variance; (3) Imbalance in the item-participant distribution in the Wright Map: low ability dead zone (-0.5 logit) and high ability ceiling effect (>1.5 logit) resulting in measurement inequity.

The theoretical implications expand Buck's model by establishing an optimal complexity threshold (~1.8 logits) and integrating Kunnan's fairness principles into the Culturally-Responsive Rasch Measurement framework. Practically, the findings recommend revising problematic items (e.g., simplifying distractors in Item 18), adding strategic items (5 easy items and 3 difficult items), and redistributing items to fill the dead zone. This study emphasizes the urgency of a psychometric approach that is sensitive to the local context to ensure the validity and fairness of language assessment in Indonesia.

REFERENCES

Abdul Aziz, A., Jusoh, M. S., Amlus, M. H., Omar, A. R., & Awang Salleh, T. S. (2014). Construct Validity: A Rasch Measurement Model Approaches. *Journal of Applied Science and Agriculture*, *9*(12), 7–12. https://www.researchgate.net/publication/266676182

Arfiani, Y., Susongko, P., & Kusuma, M. (2023). *Construct validity analysis with messick validity approach and rasch model application on scientific reasoning test items*. *6*(1), 90–105.

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing*, *38*(1), 6–40. https://doi.org/10.1177/0265532220927487

Boone, W. J., Staver, J. R., Yale, M. S., & Analysis, R. (2020). *William J . Boone , John R . Staver and Melissa S . Yale . Rasch Analysis in the Human Sciences . The Netherlands : Springer , 2014 . August 2019*. https://doi.org/10.1558/jrds.37535

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

Buck, G. (2001). Assessing Listening. In *Cambridge Language Assessment*. Cambridge University Press. https://doi.org/DOI: 10.1017/CBO9780511732959

Chapelle, C. A. (2022). Argument-Based Validation in Testing and Assessment. In

Argument-Based Validation in Testing and Assessment. https://doi.org/10.4135/9781071878811

Chapelle, C., & Lee, H. (2021). *Understanding Argument-Based Validity in Language Testing* (pp. 19–44). https://doi.org/10.1017/9781108669849.004

Dewi, H. H., Damio, S. M., & Sukarno, S. (2023). Item analysis of reading comprehension questions for English proficiency test using Rasch model. *REID (Research and Evaluation in Education)*, *9*(1), 24–36. https://doi.org/10.21831/reid.v9i1.53514

Effatpanah, F., Baghaei, P., Ravand, H., & Kunina-Habenicht, O. (2024). 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. https://doi.org/10.1080/15305058.2024.2414423

Elhambakhsh, E. (2024). *The Role of Construct Validity in Designing English Language Assessment Tasks*. *16*(34), 55–78. https://doi.org/10.22034/elt.2024.61423.2638

ETS. (2023). *toefl-ibt-test-score-data-summary-2023.pdf*.

Fan, J., & Knoch, U. (2019). *Fairness in language assessment : What can the Rasch model offer ? 8*(2), 117–142.

Forero, J., Vette, A. H., & Hebert, J. S. (2023). Technology - based balance performance assessment can eliminate floor and ceiling effects. *Scientific Reports*, *0123456789*, 1–11. https://doi.org/10.1038/s41598-023-41671-8

Futri, V. I., Rosnawati, R., Rahim, A., & Marlina, M. (2022). Rasch Model Study on Mathematics Examination Test Using Item Response Theory Approach. *International Journal on Emerging Mathematics Education*, *6*(1), 29. https://doi.org/10.12928/ijeme.v6i1.21761

Gilakjani, A. P., & Sabouri, N. B. (2016). Learners' Listening Comprehension Difficulties in English Language Learning: A Literature Review. *English Language Teaching*, *9*(6), 123. https://doi.org/10.5539/elt.v9n6p123

Goh, C.C.M., & Vandergrift, L. (2021). *Teaching and Learning Second Language Listening: Metacognition in Action (2nd ed.)*. Routledge. https://doi.org/https://doi.org/10.4324/9780429287749

Habibi, H., Jumadi, J., & Mundilarto, M. (2019). The rasch-rating scale model to identify learning difficulties of physics students based on self-regulation skills. *International*

Journal of Evaluation and Research in Education, *8*(4), 659–665. https://doi.org/10.11591/ijere.v8i4.20292

Id, D. Y. (2023). *Examining the subjective fairness of at-home and online tests : Taking Duolingo English Test as an example*. *19*, 1–19. https://doi.org/10.1371/journal.pone.0291629

Irawan, S., & Ahmad, Y. B. (2021). *Students ' Perceptions of Listening Learning Using the Bottom-up Strategy*. *4778*, 94–102. https://doi.org/10.24256/ideas.v9i2.1993

Kiran, A. (2023). English Language Assessment: Innovations, Validity, And Reliability. *Journal of International English Research Studies*, *1*(2), 1–8. https://languagejournals.com/index.php/englishjournal/article/view/8

Kirkpatrick, A. (2014). English in Southeast Asia: Pedagogical and policy implications. *World Englishes*, *33*(4), 426–438. https://doi.org/10.1111/weng.12105

Kunnan, A. J. (2010). *Statistical analyses for test fairness*. *1*.

Li, S. (2016). The Construct Validity Of Language Aptitude: A Meta-Analysis. *Studies in Second Language Acquisition*, *38*(4), 801–842. https://doi.org/DOI: 10.1017/S027226311500042X

Lia, R. M., Rusilowati, A., & Isnaeni, W. (2020). NGSS-oriented chemistry test instruments: Validity and reliability analysis with the Rasch model. *REID (Research and Evaluation in Education)*, *6*(1), 41–50. https://doi.org/10.21831/reid.v6i1.30112

Linacre, J. M. (2020). *Rasch measurement training seminars: Winsteps and Facets* (pp. 1–22).

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, *22*(3), 276–282.

Miao, Y. (2024). FACTORS AFFECTING LISTENER PERCEPTION OF ACCENTED SPEECH: THE ROLE OF ACCENT FAMILIARITY AND LINGUISTIC TRAINING. *International Journal of Listening*, *38*(3), 203–215. https://doi.org/10.1080/10904018.2023.2252019

Muchlisin, M., Mardapi, D., & Setiawati, F. A. (2019). An analysis of Javanese language test characteristic using the Rasch model in R program. *REID (Research and Evaluation in Education)*, *5*(1), 61–74. https://doi.org/10.21831/reid.v5i1.23773

Mufrihah, A. (2025). *Rasch Model Analysis of Santri Reverence Morals Scale*. *8*, 1–18.

Nishizawa, Hitoshi. (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

O'Loughlin, K. (2013). Developing the assessment literacy of university proficiency test users. *Language Testing*, *30*(3), 363–380. https://doi.org/10.1177/0265532213480336

Pinto, J. O., Dores, A. R., Peixoto, B., & Barbosa, F. (2025). Ecological validity in neurocognitive assessment: Systematized review, content analysis, and proposal of an instrument. *Applied Neuropsychology. Adult*, *32*(2), 577–594. https://doi.org/10.1080/23279095.2023.2170800

Ramadhianti, A., & Somba, S. (2022). Listening Comprehension Difficulties in Indonesian EFL Students. *Journal of Learning and Instructional Studies*, *1*(3 SE-Articles), 111–121. https://doi.org/10.46637/jlis.v1i3.7

Shaw, A. (2023). Idea-Sharing Crafting Item Difficulty in TOEFL iBT Listening Tests. *Pasaa*, *66*(October), 212–225. https://doi.org/10.58837/chula.pasaa.66.1.7

Shin, J., Guo, Q., & Gierl, M. J. (2019). *Multiple-Choice Item Distractor Development Using Topic Modeling Approaches*. *10*(April), 1–14. https://doi.org/10.3389/fpsyg.2019.00825

Vandergrift, L., & Goh, C. (2009). *The Handbook of Language Teaching*. 395–411.

Yudkowsky, R., Park, Y. S., & Downing, S. M. (2020). *Assessment in Health Professions Education*. Routledge.

Zhai, X., Haudek, K. C., Wilson, C., & Stuhlsatz, M. (2021). *A Framework of Construct-Irrelevant Variance for Contextualized Constructed Response Assessment*. *6*(October), 1–13. https://doi.org/10.3389/feduc.2021.751283