SMART
(Journal of English Language Teaching and Applied Linguistics)
Volume 11, No. 2, August 2025 Page. 176-191
p-ISSN: 2356-2048 e-ISSN: 2356-203x

RASCH MODEL-BASED EVALUATION OF TOEFL LISTENING ITEMS: ANALYZING DIFFICULTY, DISCRIMINATION, AND FIT

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

Abstract

This study analyzed TOEFL Listening Section items using Rasch Model. Quantitative analysis of 200 participants revealed significant problems: (1) 7 misfit items (Infit MNSQ >1.5), (2) 4 items with negative discrimination, 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*: Discrimination, Item Difficulty, Listening, Rasch*

INTRODUCTION

Amid growing demands for English language proficiency as a prerequisite for academic and global career success, standardized tests such as TOEFL are crucial in determining access to higher education and cross-border employment opportunities (O'Loughlin, 2013). However, a persistent deficit in listening section scores compared to other skills suggests potential issues in test item construction or cultural suitability for specific test-taker populations (Nishizawa, 2023). This disparity necessitates a critical re-evaluation of established testing instruments.

Consequently, sophisticated psychometric approaches, specifically the Rasch model within Item Response Theory (IRT), are increasingly recognized as vital for addressing the complexities of modern language assessment (C. A. Chapelle, 2022). The Rasch model offers precise analysis by modelling the interaction between item characteristics and test-taker ability, surpassing classical methods (Boone et al., 2020). Its efficacy in identifying

misfitting items and ensuring test validity is well-documented across diverse domains, including TOEFL reading comprehension Dewi et al. (2023), science Futri et al. (2022), regional languages (Muchlisin et al., 2019), and physics (Habibi et al., 2019). Nevertheless, the application of the Rasch model for comprehensive, multidimensional analysis of the TOEFL Listening Section remains relatively limited.

Furthermore, the TOEFL listening test presents unique methodological challenges due to the intricate cognitive processes involved in sound perception, meaning processing, and information integration under time constraints (Goh & Vandergrift, 2021). Recent research indicates non-linguistic factors, such as accent familiarity and speech rate, substantially influence scores for test-takers from English as a Foreign Language (EFL) contexts, challenging the presumed cultural neutrality of test items and underscoring the need for analytical approaches sensitive to these latent variables (Miao, 2024). These findings question the validity of test items previously considered culturally neutral while highlighting the need for a more analytical approach 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 educational system variations (Kirkpatrick, 2014). Indonesian test-takers, for instance, encounter specific difficulties with features prevalent in the TOEFL listening section, such as American English-connected speech and idioms (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 most significant contributors of test-takers globally. The gap between practical needs and the existing knowledge base urgently requires further investigation.

The practical implications are evident at institutions like Kusuma Husada University at Surakarta, where a disconnect exists between students meeting minimum TOEFL-like scores for graduation (65%) and those demonstrating functional academic English fluency (40%). 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 the micro-level analysis of test item quality.

Theoretically, evolving validity concepts in language testing demand a more holistic, contextual understanding, including empirical evidence on how measurement occurs across different groups (Kiran, 2023; Elhambakhsh, 2024). This paradigm shift has not been fully integrated into TOEFL listening research using modern frameworks like argument-based validity. This theoretical and practical disconnect risks perpetuating assessment tools that fail to capture authentic linguistic competencies in diverse populations. Methodologically, prior studies on TOEFL item quality often exhibit fragmentation, focusing narrowly on single parameters like difficulty or validity without integrated analysis (Li, 2016; Shaw 2023), leaving interactions between key parameters unexplored. Such compartmentalization overlooks the synergistic nature of psychometric properties essential for robust test design. The Indonesian context adds further complexity, requiring analytical sensitivity to local variables often overlooked in Western or East Asian-dominated research (Widodo & Perfecto, 2022). Neglecting these sociolinguistic nuances may inadvertently privilege certain learner profiles while disadvantaging others.

Therefore, this study aims to conduct a comprehensive Rasch model analysis of TOEFL Listening Section item quality, integrating assessment of difficulty, discrimination, and item fit. Theoretically, it seeks to enrich discussions on listening test construct validity in EFL contexts. Practically, it will provide actionable insights for test developers and institutions like Kusuma Husada University by adopting an integrative analytical approach, drawing on cross-domain principles (Habibi et al., 2019; Muchlisin et al., 2019). This research bridges traditionally separate aspects of language testing analysis.

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 (Boone et al., 2020). 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).

Participants,

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.

Instruments

Primary data were obtained from 50 validated TOEFL ITP (Form 7–9) questions (κ = 0.85) using CELP criteria (McHugh, 2012), while secondary data included historical ETS scores.

Data Analysis

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

Quantitative Analysis of Item Parameters

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. Sample Characteristics and Test Composition

Listening Type	Question Type	Number of Questions	Item Number
I. Short Dialogue (Question 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 (Question 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 (Question 39-50)	Gist/Main Idea	1	39
	Term Definition	1	40
	Detail/Process	6	41, 42, 43, 45, 46, 49
	Classification	2	47, 48
	Location/Origin	2	44, 50

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 (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 (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)

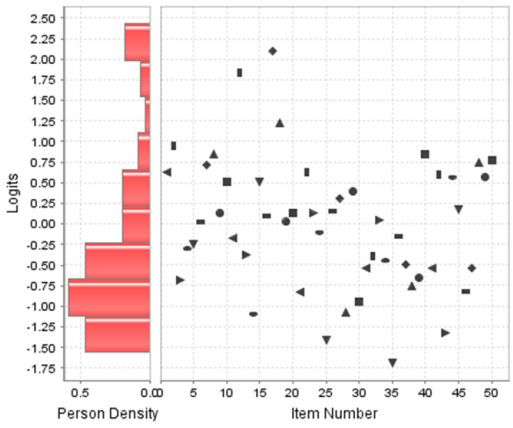
The impact of these problematic items on test reliability can be observed through *Reliability If Item Deleted* analysis. The removal of items with negative discrimination, particularly Item 18, resulted in a significant increase in Cronbach's alpha coefficient, from the original value of 0.9052 to 0.9098. 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.



Picture 1. Wright Map

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.

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 responds to the urgent need for locally contextualized psychometric approaches in Indonesia, directly addressing validity and fairness gaps in language assessment practices.

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. *Thabiea : Journal of Natural Science Teaching*, *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). Rasch Analysis in the Human Sciences. *Journal of Research Design and Statistics and Communicatiob Science, August 2019*. https://doi.org/10.1558/jrds.37535

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 *Sage Research Methods*. 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

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

Fan, J., & Knoch, U. (2019). Fairness in language assessment : What can the Rasch model

offer ?. *Language Testing and Assessment*, 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 Testas an example. PLoS ONE 18(9): e0291629*. https://doi.org/10.1371/journal. pone.0291629

Irawan, S., & Ahmad, Y. B. (2021). Students ' Perceptions of Listening Learning Using the Bottom-up Strategy. *IDEAS Journal of Language Teaching and Learning, Linguistics and Literature*, *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. *Rev. Franç. de Linguistique Appliquée*, *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

Linacre, J. M. (2020). Rasch measurement training seminars: Winsteps and Facets. *University of Sydney Australia* (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. *Islamic Guidance and Counseling Journal*, *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. *Frontiers in Psychology*, *10*(April), 1–14. https://doi.org/10.3389/fpsyg.2019.00825

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

Yudkowsky, R., Park, Y. S., & Downing, S. M. (2020). Assessment in Health Professions Education (Routledge (ed.); Sedonf Edi). Routledge.

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