Educational Leadership and Management Journal (Element)
Volume 04 Number 01 (2026)
https://online-journal.unja.ac.id/element
DOI : 10.22437/element.v4i1.56664

e-ISSN 3031-4721

141-159

Assessing the Measurement Quality of a Model Lecturer TPACK Instrument: Evidence from Rasch Analysis

Nova Susanti[1*], Dewi Iriani[2], Mia Aina [3]
[1] Physics Education, Universitas Jambi, Indonesia
[2] Mathematics Education, Universitas Jambi, Indonesia
[3] Biology Education, Universitas Jambi, Indonesia

Article Info

Article history:
Received 15, 02, 2026
Revised 01, 04, 2026
Accepted 05, 06, 2026

Keywords:
TPACK, Rasch Measurement Theory, Instrument Validation, Physics Education, Psychometric Evaluation, Lesson Study.

ABSTRACT

Technological Pedagogical Content Knowledge (TPACK) is a fundamental competency for educators in integrating technology, pedagogy, and content knowledge into effective teaching practices. Ensuring the psychometric quality of TPACK instruments is therefore essential for obtaining accurate and reliable measurement results. This study aimed to evaluate the measurement quality of a TPACK instrument for model lecturers using Rasch Measurement Theory. A quantitative survey design was employed involving 32 respondents who completed a 25-item TPACK questionnaire using a four-point Likert scale. Data were analyzed using WINSTEPS to examine respondent reliability, item reliability, item fit, item difficulty, and rating scale functioning. The results indicated excellent respondent reliability (0.91) and person separation (3.24), demonstrating that the instrument effectively differentiated respondents across multiple competency levels. However, item reliability (0.45) and item separation (0.91) revealed limited variation in item difficulty. The item difficulty measures ranged from −0.68 to 1.43 logits, with PCK4 identified as the most difficult item and TK1 and TK4 as the easiest. Most items fit the Rasch model, although several items (CK1, CK2, PK3, PCK4, and TPACK1) exhibited misfit characteristics requiring revision. Rating scale analysis showed that the four-point Likert categories generally functioned as intended, but response concentrations in the upper categories suggested a potential ceiling effect and limited discrimination among higher-ability respondents. The novelty of this study lies in the comprehensive application of Rasch Measurement Theory to evaluate a TPACK instrument within a Lesson Study context. The findings provide empirical evidence for improving instrument quality through item refinement, broader difficulty calibration, and optimization of response categories. This study contributes to the development of more valid and reliable assessment tools for measuring technology integration competencies in physics education and professional development programs.

CC BY SA

Corresponding Author: nova_fisikaunja@unja.ac.id

INTRODUCTION

The rapid advancement of digital technologies in the twenty-first century has transformed educational practices and increased expectations for educators to integrate technology effectively into teaching and learning. In higher education, lecturers are no longer required solely to master disciplinary content but also to possess the pedagogical

and technological competencies necessary to facilitate meaningful learning experiences. This competence is conceptualized as Technological Pedagogical Content Knowledge (TPACK), a framework developed by Mishra and Koehler (2006) that extends Shulman's Pedagogical Content Knowledge (PCK) model by incorporating technological knowledge into the interaction between pedagogy and content. The TPACK framework has become one of the most widely adopted models for evaluating teachers' and lecturers' readiness to implement technology-enhanced instruction in various educational contexts.

Within physics education, TPACK is particularly important because physics concepts are often abstract, mathematically intensive, and require multiple forms of representation for effective understanding. Physics learning frequently involves the integration of equations, graphs, simulations, visualizations, and experimental activities that demand sophisticated instructional design. Studies have shown that students' conceptual understanding improves when educators effectively combine pedagogical strategies with digital technologies and multimodal representations (Chai et al., 2022; Schmid et al., 2021). Furthermore, technology-supported learning environments can enhance engagement, scientific reasoning, and problem-solving skills when instructors possess adequate TPACK competencies (Koehler et al., 2024). Consequently, evaluating lecturers' TPACK competencies has become a critical issue in improving the quality of physics teaching and learning.

Recent research has extensively examined the relationships among TPACK dimensions and their influence on teaching effectiveness. Structural analyses using approaches such as Structural Equation Modeling (SEM) and Partial Least Squares (PLS-SEM) have consistently demonstrated significant relationships among technological, pedagogical, and content knowledge constructs (Valtonen et al., 2023). Although these studies provide evidence of construct relationships, they often focus on structural validity rather than the psychometric quality of individual measurement items. As a result, an instrument may exhibit acceptable construct validity while still containing poorly functioning items that reduce measurement accuracy and introduce bias.

The quality of educational measurement instruments is fundamental to producing valid and reliable assessments of professional competence. Instruments with weak psychometric properties may generate inaccurate estimates of respondents' abilities, leading to inappropriate educational decisions. Traditionally, many educational studies have relied on Classical Test Theory (CTT) to evaluate instrument quality. However, CTT statistics are highly dependent on sample characteristics and provide limited information regarding individual item performance (Bond et al., 2021). These limitations have encouraged researchers to adopt Item Response Theory (IRT) approaches, particularly the Rasch Measurement Model, which offers a more robust framework for evaluating measurement quality.

The Rasch Measurement Model, originally developed by Georg Rasch, provides probabilistic estimates that place respondent ability and item difficulty on the same interval scale. Unlike CTT, Rasch analysis enables detailed examination of item functioning, instrument reliability, dimensionality, and measurement invariance (Boone et al., 2014; Bond et al., 2021). Through Rasch analysis, researchers can evaluate person

reliability, item reliability, separation indices, item fit statistics, Wright maps, and Differential Item Functioning (DIF). These indicators provide comprehensive evidence regarding whether an instrument consistently measures a single latent construct and whether individual items function appropriately across different respondent groups.

The application of Rasch analysis is particularly relevant in evaluating TPACK instruments because TPACK represents a multidimensional and complex construct. In Lesson Study contexts, where lecturers collaboratively plan, implement, observe, and reflect on instructional practices, accurate assessment of TPACK competencies is essential for supporting professional development and evidence-based instructional improvement. The results of TPACK assessments are often used to inform training programs, teaching evaluations, and educational policy decisions. Therefore, ensuring the psychometric quality of TPACK instruments is crucial for obtaining trustworthy measurement outcomes. Despite the growing body of research on TPACK, studies employing Rasch Measurement Theory to evaluate TPACK instruments remain relatively limited, especially in the context of physics education in Indonesia. Most existing studies emphasize the relationships among TPACK variables rather than investigating the measurement characteristics of individual items. Consequently, there remains a significant research gap regarding the reliability, validity, and item functioning of TPACK instruments designed for model lecturers engaged in Lesson Study activities. Addressing this gap is important because high-quality instruments are necessary to accurately assess lecturers' competencies and support the implementation of technology-enhanced physics instruction.

Therefore, this study aims to evaluate the psychometric quality of a TPACK instrument for model lecturers using the Rasch Measurement Model. Specifically, the study examines construct validity, reliability, item fit, dimensionality, differential item functioning, and the distribution of respondent ability and item difficulty. The novelty of this study lies in the comprehensive application of Rasch Measurement Theory to assess a TPACK instrument within a Lesson Study-based physics education context. The findings are expected to contribute to the advancement of physics education research by providing empirical evidence for the development of valid, reliable, and objective instruments for assessing lecturers' professional competencies in technology-integrated teaching environments.

LITERATUR REVIEW

A review of studies published between 2015 and 2025 demonstrates growing scholarly interest in Technological Pedagogical Content Knowledge (TPACK) as a framework for understanding educators' readiness to integrate technology into teaching and learning. Since its introduction by Mishra and Koehler (2006), TPACK has become one of the most widely adopted frameworks for examining how technological, pedagogical, and content knowledge interact to support effective instruction. Recent developments in digital learning, artificial intelligence, and educational technology have further strengthened the relevance of TPACK as a foundation for teacher and lecturer professional competence in technology-rich educational environments (Mishra et al., 2023; Tseng et al., 2022).

The increasing use of digital technologies in education has generated a strong demand for valid and reliable instruments capable of assessing TPACK competencies. Educational institutions are expected not only to provide technological infrastructure but also to ensure that educators possess the knowledge required to integrate technology meaningfully into instructional practice. Consequently, numerous studies have focused on developing and validating TPACK instruments across various educational contexts, including teacher education, language education, science education, and higher education settings. These studies consistently demonstrate that TPACK plays an important role in supporting instructional innovation, student engagement, and technology-enhanced learning outcomes (Ammade et al., 2020; Chai et al., 2022).

Despite the growing body of literature, considerable debate remains regarding the most appropriate methods for assessing TPACK. Many studies rely on self-report questionnaires analyzed through Classical Test Theory (CTT), Confirmatory Factor Analysis (CFA), or Structural Equation Modeling (SEM). These approaches have been useful for examining latent constructs and validating theoretical relationships among TPACK dimensions. However, scholars have increasingly questioned whether these methods provide sufficient evidence regarding item-level performance and measurement precision. In particular, they often offer limited information about how individual items function across varying respondent ability levels and whether response patterns are consistent with measurement assumptions (Kadluba et al., 2025; Zhao & Wang, 2024).

As a result, recent TPACK research has increasingly incorporated modern psychometric approaches such as Rasch Measurement Theory. Unlike traditional approaches that primarily focus on scale-level statistics, Rasch analysis evaluates the interaction between respondents and items on a common measurement continuum. This approach enables researchers to examine item difficulty, respondent ability, item fit, rating scale functioning, dimensionality, and measurement invariance. Such information is particularly valuable because TPACK is a complex construct involving multiple domains of knowledge that may not be captured adequately through total scores alone (Saubern & Urbach, 2024; Saubern, 2025).

A critical analysis of the literature reveals several important research gaps. First, most TPACK validation studies have been conducted in general education, teacher preparation programs, mathematics education, and language education. Comparatively fewer studies have focused specifically on physics education contexts, despite the fact that technology integration in physics teaching often requires specialized competencies involving simulations, virtual laboratories, computational modeling, and scientific visualization tools. Because technology use in physics education differs substantially from other disciplines, findings from general educational settings may not be directly transferable to physics education environments (Kadluba et al., 2025; Tseng et al., 2022).

Second, previous studies have primarily emphasized structural validation and construct relationships rather than detailed psychometric evaluation. Although CFA and SEM provide valuable evidence regarding factor structures and theoretical models, they often provide limited insight into item functioning, item hierarchy, response category performance, and measurement accuracy. Consequently, many TPACK instruments

demonstrate acceptable reliability at the scale level while offering relatively little evidence regarding the quality of individual items. This limitation creates uncertainty regarding whether instrument scores accurately represent respondents' actual TPACK competencies. Researchers have therefore recommended greater use of Rasch-based validation approaches to strengthen evidence of construct validity and measurement quality (Saubern & Urbach, 2024; Ismail et al., 2025).

The third gap concerns the population under investigation. Existing studies have predominantly focused on school teachers, pre-service teachers, or lecturers in general higher education settings. Limited attention has been directed toward model lecturers involved in Lesson Study activities. This represents an important omission because Lesson Study emphasizes collaborative planning, classroom observation, reflective practice, and continuous instructional improvement. Such activities provide unique opportunities for integrating technological, pedagogical, and content knowledge in authentic teaching situations. Furthermore, Lesson Study encourages collective professional learning, which aligns closely with recent perspectives that view TPACK not only as an individual competency but also as a collaborative and socially constructed form of professional knowledge (Yeh et al., 2021).

The importance of examining TPACK within Lesson Study contexts has become increasingly evident as higher education institutions seek to strengthen technology-enhanced teaching practices. Model lecturers participating in Lesson Study often function as instructional leaders who influence curriculum implementation, pedagogical innovation, and professional development within their departments. Their ability to integrate technology effectively may therefore have broader implications for institutional teaching quality. However, empirical evidence regarding the measurement of TPACK among this population remains limited, particularly within physics education programs. This gap highlights the need for context-specific validation studies capable of providing accurate assessments of lecturers' technology integration competencies (Yeh et al., 2021; Mishra et al., 2023).

To address these gaps, the present study employs Rasch Measurement Theory to conduct a comprehensive psychometric evaluation of a TPACK instrument designed for model lecturers in a Physics Education Study Program. Unlike many previous investigations that focused primarily on structural relationships among TPACK dimensions, this study examines multiple indicators of measurement quality, including item reliability, person reliability, separation indices, item fit statistics, dimensionality, Wright maps, rating scale functioning, and Differential Item Functioning (DIF). These analyses provide a more detailed understanding of how the instrument functions and whether it accurately measures lecturers' competencies in technology-integrated teaching environments (Aryadoust et al., 2021; Saubern, 2025).

The study contributes both methodologically and substantively to the TPACK literature. Methodologically, it expands the use of Rasch Measurement Theory in TPACK assessment by providing evidence at the item level rather than relying solely on scale-level validation procedures. Substantively, it contributes to physics education research by examining a population and educational context that have received limited attention in previous studies. The findings are expected to support the development of more accurate,

objective, and meaningful assessment instruments that can inform professional development initiatives and strengthen technology-enhanced teaching practices. Ultimately, the study responds to current calls for stronger psychometric evidence in TPACK research and provides a foundation for improving the assessment of educators' competencies in increasingly complex digital learning environments (Ismail et al., 2025; Mishra et al., 2023).

METHODS

Research Design

This study employed a descriptive quantitative survey design to evaluate the psychometric quality of a Technological Pedagogical Content Knowledge (TPACK) instrument using Rasch Measurement Theory. The Rasch approach was selected because it provides a more robust measurement framework than raw-score analysis by transforming ordinal responses into interval-level measures and simultaneously estimating respondent ability and item difficulty on a common logit scale (Bond et al., 2021). Furthermore, Rasch analysis enables researchers to examine item functioning, construct validity, reliability, dimensionality, and measurement invariance, thereby providing stronger evidence for instrument quality than conventional Classical Test Theory approaches (Boone et al., 2014; Linacre, 2023).

Research Participants

The study was conducted in the Physics Education Study Program, Faculty of Teacher Training and Education, Universitas Jambi, Indonesia, during the 2025–2026 academic year. The target population consisted of individuals directly involved in Lesson Study activities, including model lecturers and observers responsible for evaluating teaching practices. A total sampling technique was employed to ensure comprehensive representation of the population. Of the 42 participants invited to complete the survey, 33 submitted complete responses and were included in the final analysis. According to Rasch measurement guidelines, a sample size of 30–50 respondents is adequate for preliminary instrument calibration and item functioning analysis (Linacre, 2023).

Research Instrument

The research instrument was a TPACK questionnaire developed based on the seven domains of the TPACK framework, namely Technological Knowledge (TK), Pedagogical Knowledge (PK), Content Knowledge (CK), Technological Pedagogical Knowledge (TPK), Technological Content Knowledge (TCK), Pedagogical Content Knowledge (PCK), and Technological Pedagogical Content Knowledge (TPACK). The instrument consisted of 25 items measured using a four-point Likert scale ranging from 1 (strongly disagree) to 4 (strongly agree). The use of a four-point scale was intended to minimize neutral responses and encourage respondents to express more definite judgments regarding their competencies. Prior to data collection, the instrument underwent expert validation involving specialists in physics education, educational technology, and educational measurement. This process ensured content relevance, conceptual clarity, and linguistic

appropriateness of the questionnaire items. Expert review is widely recommended as an essential procedure for establishing content validity before psychometric evaluation (Boateng et al., 2018).

Data Collection Procedure

Data were collected using an online questionnaire distributed through Google Forms after the completion of Lesson Study activities. Participants were informed about the purpose of the study and voluntarily completed the survey. To ensure data quality, responses were screened for completeness before analysis. Incomplete questionnaires were excluded from the dataset. The data collection process followed four stages: (1) instrument development based on the TPACK framework, (2) expert validation and revision, (3) questionnaire administration, and (4) psychometric evaluation using Rasch analysis.

Data Analysis

Data were analyzed using WINSTEPS version 5.10.3.0 based on Rasch Measurement Theory. Several psychometric indicators were examined to determine the quality of the instrument. Reliability analysis was conducted by examining person reliability and item reliability coefficients. These indicators describe the consistency of respondent responses and the stability of item calibration. Separation indices were also calculated to evaluate the instrument's ability to distinguish respondents with different competency levels and to classify item difficulty into meaningful strata (Bond et al., 2021).

Item fit analysis was performed using Outfit Mean Square (MNSQ), Infit Mean Square (MNSQ), standardized Z-values (ZSTD), and Point-Measure Correlation (PTMEA Corr). Following established Rasch criteria, items were considered acceptable when MNSQ values ranged between 0.50 and 1.50, ZSTD values fell between –2.0 and +2.0, and PTMEA Corr values were positive and within an acceptable range (Boone et al., 2014; Linacre, 2023). Items that failed to meet these criteria were identified as potential misfit items and considered for revision.

Construct validity was evaluated through Principal Component Analysis of Residuals (PCAR) to examine the unidimensionality assumption. An instrument was considered sufficiently unidimensional when the variance explained by the Rasch measures exceeded the minimum recommended threshold and the unexplained variance in the first contrast remained within acceptable limits (Bond et al., 2021).

A Wright Person–Item Map was generated to visualize the distribution of respondent abilities and item difficulties on the same logit scale. This analysis provides valuable information regarding the alignment between item difficulty and respondent competency levels, thereby identifying whether the instrument adequately targets the intended population.

Additionally, Differential Item Functioning (DIF) analysis was conducted to detect potential item bias across respondent groups whenever demographic comparison variables were available. DIF analysis helps ensure measurement fairness by identifying items that function differently across groups despite equivalent levels of the underlying construct (Boone et al., 2014).

Ethical Considerations

Participation in this study was voluntary, and all respondents provided informed consent prior to completing the questionnaire. To protect participant confidentiality, all responses were anonymized before analysis. The collected data were used exclusively for academic research purposes and were stored securely throughout the research process. Ethical procedures followed accepted standards for educational research involving human participants.

FINDINGS AND DISCUSSION

Table 1 presents the Rasch item statistics for the TPACK instrument. The analysis was performed to examine the quality of each item in terms of difficulty level, model fit, discrimination, and response consistency. Item difficulty was evaluated using logit measures, while item fit was assessed through Infit and Outfit Mean Square (MNSQ) statistics. In addition, Point-Measure Correlation (PTMEA Corr.) and Exact Match indices were examined to determine whether each item functioned consistently with the underlying TPACK construct. According to Rasch guidelines, items with MNSQ values between 0.50 and 1.50 and positive PTMEA correlations are generally considered productive for measurement (Linacre, 2023). The results provide evidence regarding the strengths and limitations of the instrument in measuring lecturers' TPACK competencies.

Tabel 1. Item Statistics For The TPACK Instrument

Item	Measure (Logit)	S.E.	INFIT		OUTFIT		PTMEA		Exact Match	
			MNSQ	ZSTD	MNSQ	ZSTD	Corr.	Exp.	Obs (%)	Exp (%)
CK1	0.69	0.36	1.68	2.10	1.75	2.11	0.53	0.73	63.3	69.4
CK2	0.81	0.36	1.57	1.80	1.63	1.86	0.55	0.73	53.3	69.9
CK3	-0.66	0.38	0.72	-1.18	0.60	-1.07	0.75	0.65	76.7	68.7
PK1	0.16	0.36	1.30	1.12	1.41	1.25	0.56	0.70	70.0	68.8
PK2	0.43	0.36	0.65	-1.35	0.62	-1.32	0.77	0.71	80.0	69.0
PK3	0.16	0.36	1.74	2.35	1.64	1.78	0.57	0.70	73.3	68.8
PK4	0.81	0.36	1.42	1.39	1.55	1.67	0.68	0.73	53.3	69.9
PK5	0.17	0.38	1.38	1.32	1.28	0.88	0.57	0.67	79.3	69.3
PK6	-0.38	0.37	0.51	-2.31	0.48	-1.69	0.77	0.67	80.0	67.7
TK1	-0.68	0.39	0.72	-1.14	0.65	-0.84	0.71	0.60	75.9	69.2

Item	Measure (Logit)	S.E.	INFIT		OUTFIT		PTMEA		Exact Match	
			MNSQ	ZSTD	MNSQ	ZSTD	Corr.	Exp.	Obs (%)	Exp (%)
TK2	-0.52	0.39	0.77	-0.89	0.71	-0.71	0.69	0.61	79.3	68.5
TK3	-0.52	0.39	0.60	-1.77	0.55	-1.27	0.71	0.61	86.2	68.5
TK4	-0.68	0.39	0.71	-1.22	0.66	-0.81	0.71	0.60	82.8	69.2
PCK1	0.27	0.37	0.97	-0.01	1.58	1.63	0.66	0.71	65.5	68.5
PCK2	0.16	0.36	0.75	-0.91	0.73	-0.82	0.77	0.70	70.0	68.8
PCK3	-0.11	0.37	0.78	-0.82	0.66	-1.05	0.77	0.68	73.3	68.2
PCK4	1.43	0.35	1.64	1.97	1.86	2.40	0.55	0.75	56.7	68.1
TCK1	-0.38	0.37	1.28	1.11	1.09	0.37	0.73	0.67	56.7	67.7
TCK2	0.03	0.37	0.66	-1.41	0.55	-1.52	0.79	0.69	73.3	68.5
TCK3	-0.66	0.38	1.04	0.23	0.90	-0.15	0.69	0.65	73.3	68.7
TCK4	0.03	0.37	0.52	-2.12	0.44	-2.07	0.78	0.69	86.7	68.5
TPK1	-0.38	0.37	0.95	-0.14	0.82	-0.40	0.76	0.67	76.7	67.7
TPK2	-0.24	0.37	0.57	-1.91	0.46	-1.83	0.79	0.68	83.3	68.0
TPK3	-0.38	0.37	0.64	-1.58	0.55	-1.39	0.76	0.67	73.3	67.7
TPACK1	0.43	0.36	1.52	1.72	1.42	1.31	0.60	0.71	70.0	69.0
Mean	**0.00**	**0.37**	**1.00**	**-0.14**	**0.98**	**-0.07**	-	**72.5**		**68.6**
SD	**0.54**	**0.01**	**0.40**	**1.49**	**0.47**	**1.41**	-	**9.5**		**0.6**

The Rasch analysis demonstrated that the TPACK instrument possesses strong respondent reliability, as indicated by a person reliability coefficient of 0.91 and a person separation index of 3.24. These values suggest that the instrument consistently differentiates respondents into several levels of TPACK competency. In Rasch measurement, person reliability reflects the consistency of response patterns, whereas person separation indicates the instrument's ability to distinguish between respondents with different ability levels. A separation index greater than 2.0 is generally considered adequate for classifying respondents into distinct strata of ability (Bond et al., 2021). The findings therefore indicate that the instrument is effective in measuring variations in

lecturers' technological, pedagogical, and content knowledge competencies. Similar results were reported in recent TPACK validation studies, where high person reliability and separation values indicated stable measurement of educators' readiness to integrate technology into teaching.

Despite the strong respondent reliability, the instrument exhibited relatively low item reliability (0.45) and item separation (0.91). These results indicate that the items do not vary substantially in difficulty and therefore provide limited information across different competency levels. Low item separation suggests that the instrument has difficulty distinguishing between easy and difficult items, reducing its measurement precision. Previous simulation studies have demonstrated that item separation reliability is strongly influenced by sample size, test length, and person–item targeting. When respondents possess relatively similar ability levels or when items cluster within a narrow difficulty range, item reliability tends to decrease. The present findings suggest that most respondents possessed relatively high TPACK competency, while the majority of items represented moderate difficulty levels, resulting in a limited spread of item calibrations.

The item measure analysis further revealed that item PCK4 (1.43 logits) was the most difficult item, whereas items TK1 and TK4 (-0.68 logits) were the easiest. This distribution indicates that respondents found certain aspects of TPACK considerably more challenging than others. However, the overall item difficulty range was relatively narrow (SD = 0.54 logits), suggesting that the instrument may not adequately capture the full continuum of TPACK competencies. Recent Rasch-based TPACK studies have emphasized the importance of including items with diverse difficulty levels to ensure comprehensive assessment of technological, pedagogical, and content integration skills.

The item fit analysis showed that most items functioned appropriately within the Rasch model, with average Infit MNSQ (1.00) and Outfit MNSQ (0.98) values falling within the recommended range of 0.50–1.50. These findings support the construct validity of the instrument and indicate that the majority of items contribute meaningfully to the measurement of TPACK competency. Nevertheless, several items, including PCK4, CK2, CK1, PK3, and TPACK1, demonstrated misfit characteristics. Misfitting items may arise when respondents interpret item statements differently, when items measure multiple constructs simultaneously, or when item wording lacks clarity. Previous Rasch validation studies have similarly identified misfit items as indicators of potential construct ambiguity and have recommended item revision rather than immediate deletion to preserve theoretical coverage.

Another important finding concerns the functioning of the four-point Likert rating scale. The category frequency analysis showed that responses were heavily concentrated in categories 3 ("Agree") and 4 ("Strongly Agree"), while categories 1 and 2 were used infrequently. Several items also exhibited non-monotonic category progression, indicated by the notation "Average ability does not ascend with category score." This pattern suggests that respondents with similar ability levels tended to select adjacent categories inconsistently. According to Rasch rating scale theory, response categories should demonstrate a monotonic increase in average measures across categories, indicating that higher categories correspond to higher levels of the latent trait. The observed category

disordering may therefore indicate that respondents had difficulty distinguishing between some response options, particularly between "Agree" and "Strongly Agree." Future revisions may consider refining category descriptors or expanding the response scale to improve discrimination among respondents.

Overall, the findings suggest that the TPACK instrument possesses strong respondent reliability and acceptable construct representation but requires refinement at the item and rating scale levels. Specifically, future development should focus on revising misfitting items, increasing the diversity of item difficulty, and improving category functioning. Such improvements would enhance measurement precision and strengthen the instrument's ability to assess TPACK competencies across different populations of educators.

The study also demonstrates the value of Rasch Measurement Theory as a comprehensive approach for evaluating educational instruments. Unlike Classical Test Theory, Rasch analysis simultaneously examines respondent ability, item difficulty, item functioning, and rating scale performance within a unified measurement framework. This capability provides deeper insights into instrument quality and supports the development of more valid and reliable assessment tools for evaluating technology integration competencies in physics education and teacher professional development.

While the preceding analyses confirmed the overall measurement quality of the instrument, the effectiveness of the response categories must also be examined to ensure accurate measurement. Therefore, a rating scale analysis was performed. As shown in Figure 1, the category functioning curve depicts the progression of observed scores across different levels of respondent ability, providing evidence regarding the appropriateness of the four-point Likert scale used in the TPACK instrument.

Rating Scale Functioning Analysis

Figure 1 presents the category functioning analysis of the four-point Likert scale used in the TPACK instrument. The graph illustrates the relationship between respondents' ability levels and the probability of selecting each response category across the latent trait continuum. In Rasch measurement, a well-functioning rating scale should demonstrate a monotonic progression in which respondents with higher ability levels are more likely to select higher response categories. Such a pattern indicates that the response options are capable of distinguishing respondents according to their level of competency and contribute positively to measurement accuracy.

The results indicate that the empirical category scores generally follow the expected Rasch model pattern. As respondent ability increases, the average observed score also increases progressively from category 1 to category 4. This finding suggests that the response categories operate in the intended direction and that respondents are able to use the scale in a manner consistent with their perceived TPACK competency levels. The ordered progression of categories provides initial evidence that the instrument is capable of measuring variations in technological, pedagogical, and content knowledge competencies among lecturers.

The observed pattern is consistent with recommendations in Rasch measurement theory, which emphasize the importance of ordered category functioning for ensuring

valid measurement (Aryadoust et al., 2021; Medvedev & Krägeloh, 2022). When categories function appropriately, respondents with lower levels of the latent trait tend to select lower response options, whereas respondents with higher levels of the trait are more likely to choose higher categories. Similar findings have been reported in TPACK validation studies conducted by Lestari and Rispatiningsih (2023) and Ismail et al. (2025), where properly ordered response categories were considered evidence that the rating scale effectively supported construct measurement and interpretation.

Despite the generally positive findings, several irregularities can be observed in the lower and middle regions of the ability continuum, particularly between categories 2 and 3. The empirical curve does not consistently follow a smooth monotonic trajectory and exhibits noticeable fluctuations around the expected model curve. This pattern suggests that respondents may have experienced some difficulty distinguishing between these adjacent response categories. In practical terms, individuals with similar levels of TPACK competence may have interpreted the meanings of categories 2 and 3 differently, resulting in inconsistent category selection.

From a Rasch measurement perspective, such fluctuations may indicate overlapping category thresholds or partial category disordering. Previous studies have shown that when adjacent categories are not clearly differentiated, respondents tend to alternate between response options that are conceptually similar, thereby reducing measurement precision (Reise et al., 2023; Petrillo et al., 2024). This condition may also reflect ambiguity in the wording of category descriptors, particularly when intermediate response options do not represent sufficiently distinct levels of agreement or competency. Consequently, the scale may be less effective in discriminating respondents with moderate levels of TPACK ability.

Nevertheless, the overall category functioning results support the continued use of the four-point Likert scale within the TPACK instrument. The categories generally operate in the expected order and provide meaningful information about respondent ability levels. However, future revisions should consider refining the descriptors of categories 2 and 3 to improve clarity and reduce overlap between adjacent response options. Such improvements could enhance category discrimination, strengthen measurement precision, and increase the psychometric quality of the instrument. Consistent with the recommendations of Saubern and Urbach (2024) and Saubern (2025), ongoing refinement of rating scale categories is essential for ensuring that TPACK instruments accurately capture competency differences across the full range of respondent abilities.

Another notable finding is the strong concentration of responses at the upper end of the rating scale. The category probability curve approached the highest score category relatively early and remained stable across higher levels of respondent ability. This pattern indicates a potential ceiling effect, suggesting that respondents with moderate to high levels of TPACK competence tended to select the highest response category. As a result, the instrument may have limited capacity to differentiate respondents who already possess relatively advanced technological, pedagogical, and content knowledge competencies. This finding is consistent with the descriptive Rasch results, which showed that the mean person measure exceeded the mean item difficulty, indicating that

respondents generally perceived the instrument items as relatively easy. Similar patterns have been reported in previous TPACK studies where participants demonstrated high levels of readiness or competence in technology integration, resulting in clustering at the upper categories of the rating scale (Ismail et al., 2025; Saubern, 2025).

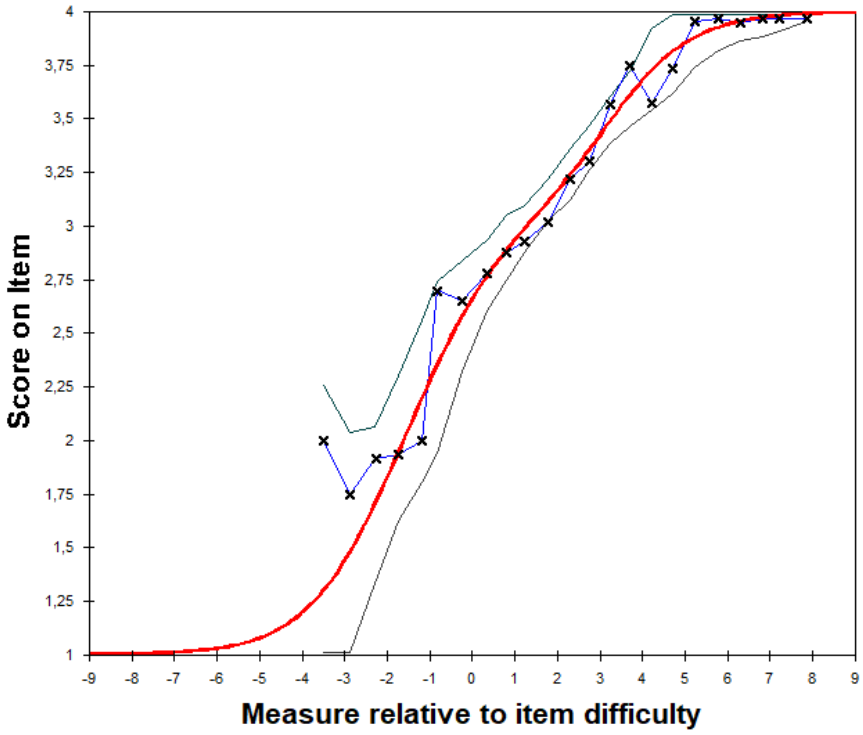

Figure 1. the category functioning analysis

From a measurement perspective, the presence of a ceiling effect does not necessarily indicate poor instrument quality; however, it highlights limitations in the instrument's ability to distinguish among respondents with higher ability levels. According to Rasch measurement theory, an effective instrument should provide sufficient item difficulty variation to accurately estimate respondents across the entire latent trait continuum (Medvedev & Krägeloh, 2022). When most respondents consistently choose the highest category, the amount of information available for differentiating high-performing individuals decreases, leading to reduced measurement precision at the upper end of the scale. Consequently, the instrument may underestimate variability among respondents who possess advanced TPACK competencies because their responses become concentrated within a narrow score range.

The observed pattern may also reflect the increasing integration of digital technologies within higher education environments. In recent years, lecturers have become more familiar with technology-enhanced teaching practices due to the rapid adoption of digital learning platforms, online instructional tools, and blended learning approaches. Such developments may contribute to generally high self-reported TPACK levels among respondents. Similar findings were reported by Islam et al. (2020), who found that higher education instructors frequently demonstrated positive perceptions of ICT integration, resulting in elevated response distributions on technology-related

assessment instruments. Therefore, the ceiling effect observed in this study may partially reflect genuine respondent competence rather than solely a measurement limitation.

Another important finding concerns the overlap observed between categories 2 and 3 on the rating scale. Ideally, each response category should represent a distinct level of the latent trait and function independently in discriminating respondent abilities. However, the category probability curves indicated that respondents may not clearly distinguish between these two adjacent categories. This suggests that the wording or conceptual meaning of the category descriptors may not be sufficiently differentiated, causing respondents with similar competency levels to alternate between categories without a clear pattern. Such overlap has important implications because category ambiguity can reduce the effectiveness of the rating scale and introduce additional measurement error.

Rasch scholars emphasize that rating scale evaluation should include examination of threshold ordering, category frequencies, and category functioning to determine whether respondents utilize response options as intended (Aryadoust et al., 2021). When adjacent categories exhibit substantial overlap, researchers are encouraged to consider revising category labels, redefining category descriptions, or, in some cases, collapsing categories to improve measurement performance. Reise et al. (2023) similarly argued that ineffective category functioning can obscure meaningful differences among respondents and reduce the interpretability of assessment results. Therefore, the overlap identified in this study suggests that future revisions should focus on improving the clarity and distinctiveness of the intermediate response categories.

The findings are also consistent with previous Rasch-based validation studies in educational measurement. Lestari and Rispatiningsih (2023), in their validation of a TPACK instrument for pre-service chemistry teachers, highlighted the importance of evaluating category functioning to ensure that respondents can distinguish between response options accurately. Likewise, Maulana et al. (2026) reported that response category performance significantly influences the precision of person ability estimation in Rasch analysis. These studies reinforce the notion that instrument validity extends beyond item content and includes the effectiveness of the response scale used to capture respondent perceptions.

In addition to category refinement, the findings suggest a need to expand the range of item difficulty within the instrument. The relatively high person measures compared to item difficulty levels indicate that the current set of items may not adequately challenge respondents with advanced TPACK competencies. According to Saubern and Urbach (2024), one of the strengths of item response theory and Rasch modeling is the ability to identify mismatches between respondent ability and item difficulty. When items are concentrated at lower or moderate difficulty levels, the instrument becomes less effective at measuring respondents who possess higher levels of competence. Therefore, introducing more complex and demanding TPACK items could enhance the instrument's ability to discriminate among highly competent lecturers.

This recommendation is supported by recent developments in TPACK assessment research. Saubern (2025) argues that contemporary TPACK measurement should move beyond basic technology integration indicators and incorporate items that capture higher-

order competencies, such as instructional innovation, pedagogical adaptation, technology-supported problem-solving, and the critical evaluation of digital learning resources. By including more sophisticated indicators, researchers can better differentiate respondents across a broader spectrum of TPACK proficiency and reduce the likelihood of ceiling effects.

Furthermore, the findings underscore the value of Rasch measurement theory as a rigorous framework for instrument validation. Unlike traditional reliability approaches that focus primarily on total scores, Rasch analysis provides detailed information regarding item functioning, person ability distribution, category performance, and measurement precision (McLaughlin et al., 2023; Petrillo et al., 2024). This level of diagnostic information enables researchers to identify specific weaknesses within an instrument and implement targeted improvements. In the present study, the identification of category overlap and potential ceiling effects illustrates how Rasch analysis can contribute to the ongoing refinement of educational assessment instruments.

Overall, the rating scale analysis provides substantial evidence supporting the use of the TPACK instrument for assessing lecturers' technological, pedagogical, and content knowledge competencies. The response categories generally functioned as intended, and the instrument demonstrated the capacity to measure the targeted construct. However, several areas for improvement remain. Future revisions should focus on enhancing category differentiation, particularly between intermediate response categories, and expanding item difficulty levels to better capture respondents with advanced TPACK competencies. Such refinements would improve measurement precision, strengthen the psychometric quality of the instrument, and increase its usefulness for evaluating technology integration competencies in physics education and other higher education contexts. Ultimately, continuous instrument development based on Rasch measurement principles will contribute to more accurate, reliable, and meaningful assessments of TPACK, thereby supporting evidence-based educational improvement and professional development initiatives.

CONCLUSION

This study evaluated the psychometric quality of a Technological Pedagogical Content Knowledge (TPACK) instrument using Rasch Measurement Theory. The findings indicate that the instrument demonstrates strong respondent reliability, as evidenced by a person reliability coefficient of 0.91 and a person separation index of 3.24, suggesting that the instrument can consistently differentiate respondents across multiple levels of TPACK competency. These results confirm that the instrument is effective in measuring variations in lecturers' technological, pedagogical, and content knowledge competencies.

However, the item-level analysis revealed several limitations. The item reliability coefficient (0.45) and item separation index (0.91) indicate a limited range of item difficulty, suggesting that the instrument is less effective in distinguishing among different levels of item complexity. Furthermore, several items (PCK4, CK2, CK1, PK3, and TPACK1) exhibited misfit characteristics, indicating the need for revision to improve construct representation and measurement precision. The rating scale analysis also revealed that the four-point Likert categories did not function optimally, as responses were

concentrated in the higher categories, indicating a potential ceiling effect and limited discrimination among respondents with moderate and high competency levels.

The novelty of this study lies in the comprehensive application of Rasch Measurement Theory to evaluate the quality of a TPACK instrument for model lecturers. Unlike conventional validation approaches, Rasch analysis provided detailed evidence regarding respondent reliability, item functioning, and category effectiveness. The findings contribute to the advancement of educational measurement by demonstrating the importance of examining both item quality and rating scale performance when assessing TPACK competencies.

Future research should focus on revising misfitting items, expanding the range of item difficulty, and refining response categories to enhance measurement accuracy. Such improvements will strengthen the instrument's psychometric properties and support more reliable assessments of technology integration competencies in physics education and professional development contexts.

ACKNOWLEDGMENTS

Acknowledgements, avoiding identifying any of the authors before peer review.

AUTHOR CONTRIBUTIONS

Author 1: Conceptualization, methodology, instrument development, data collection, Rasch analysis, interpretation of results, writing—original draft preparation, and manuscript revision; Author 2: Data curation, instrument validation, formal analysis, visualization, writing—review and editing, and manuscript refinement; Author 3: Supervision, methodological validation, interpretation of findings, critical review of the manuscript, and final approval of the submitted version.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Ammade, S., Mahmud, M., Jabu, B., & Tahmir, S. (2020). TPACK model based instruction in teaching writing: An analysis on TPACK literacy. *International Journal of Language Education, 4*(1), 129–140. https://doi.org/10.26858/ijole.v4i2.12465

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

Asrijanty. (2025). Applying the Rasch Model as a Diagnostic Tool for Rating Scale Refinement. *JP3I: Jurnal Pengukuran Psikologi dan Pendidikan Indonesia*, *15*(1). https://doi.org/10.15408/jp3i.v15i1.50616

Bintang, R. S., & Suprananto. (2024). The impact of sample size, test length, and person-item targeting on the separation reliability in Rasch model: A simulation study. *JP3I:*

Jurnal Pengukuran Psikologi dan Pendidikan Indonesia, *13*(1), 1–13. https://doi.org/10.15408/jp3i.v13i1.27975

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research. *Frontiers in Public Health*, *6*, Article 149. https://doi.org/10.3389/fpubh.2018.00149

Bond, T. G., Yan, Z., & Heene, M. (2021). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (4th ed.). Routledge. https://doi.org/10.4324/9780429030499

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. Springer. https://doi.org/10.1558/jrds.37535

Chaeruman, U. A., Suwarno, S., Maudiarti, S., & Sudatha, I. G. W. (2025). Measuring TPACK in context: Validating a cross-contextually fair instrument for Indonesian teachers. *Indonesian Journal of Educational Development*. https://doi.org/10.59672/ijed.v7i1.6354

Chai, C. S., Jong, M. S. Y., Yin, H., & Tan, L. (2022). Validating and modelling technological pedagogical content knowledge framework among educators: A review of recent developments. *Educational Technology & Society*, *25*(2), 1–15. https://doi.org/10.14742/ajet.174

Chong, J., Mokshein, S. E., & Mustapha, R. (2022). Applying the Rasch Rating Scale Model (RSM) to investigate the rating scales function in survey research instrument. *Cakrawala Pendidikan*, *41*(1), 97–111. https://doi.org/10.21831/cp.v41i1.39130

Fadzil, A. H. P., Maat, S. M., & Mahmud, M. S. (2022). A Rasch model analysis of the TPACK instrument in the creative teaching of primary mathematics teachers. *Cypriot Journal of Educational Sciences*, *17*(11), 4259–4274. https://doi.org/10.18844/cjes.v17i11.7792

Islam, A. A., Gu, X., Crook, C., & Spector, J. M. (2020). Assessment of ICT in tertiary education applying structural equation modeling and Rasch model. *SAGE Open, 10*(4), 2158244020975409. https://doi.org/10.1177/2158244020975409

Ismail, I., Riandi, R., Kaniawati, I., et al. (2025). Assessing science teachers' readiness for technology-integrated green energy instruction: Development and validation of TPACK instrument using Rasch analysis. *Social Sciences & Humanities Open*, *12*, Article 101948. https://doi.org/10.1016/j.ssaho.2025.101948

Ismail, I., Riandi, R., Kaniawati, I., Sopandi, W., Soeharto, S., Rochman, S., Supriyadi, S., et al. (2025). Assessing science teachers' readiness for technology-integrated green energy instruction: Development and validation of TPACK instrument using Rasch analysis. *Social Sciences & Humanities Open, 12*, 101948. https://doi.org/10.1016/j.ssaho.2025.101948

Kadluba, A., Strohmaier, A., Schons, C., & Obersteiner, A. (2025). How much C is in TPACK? A systematic review on the assessment of TPACK in mathematics. *Educational Studies in Mathematics, 118*(2), 169–199. https://doi.org/10.1007/s10649-024-10327-4

Koehler, M. J., Mishra, P., Kereluik, K., Shin, T. S., & Graham, C. R. (2024). The technological pedagogical content knowledge framework: Current perspectives and

future directions. *Computers & Education*, *210*, Article 104998. https://link.springer.com/chapter/10.1007/978-1-4614-3185-5_9

Lestari, W. T., & Rispatiningsih, D. M. (2023). Validitas isi instrumen TPACK pada calon guru kimia menggunakan aplikasi model Rasch. *JIMAD: Jurnal Ilmiah Mutiara Pendidikan, 1*(2), 16–25.

Linacre, J. M. (2023). *A User's Guide to Winsteps Rasch-Model Computer Programs*. Winsteps.com.

Maulana, I., Syarah, E., Nainggolan, E., & Qudratuddarsi, H. (2026). Validation of instrument to measure pre-service science teachers digital skills: Rasch model. *Jurnal Ilmiah Profesi Pendidikan, 11*(2), 1987–1996.

McLaughlin, J. E., Angelo, T. A., & White, P. J. (2023). Validating criteria for identifying core concepts using many-facet Rasch measurement. *Frontiers in Education, 8*, Article 1150781. https://doi.org/10.3389/feduc.2023.1150781

Medvedev, O. N., & Krägeloh, C. U. (2022). Rasch measurement model. In R. A. Baer (Ed.), *Handbook of assessment in mindfulness research* (pp. 1–18). Springer. https://doi.org/10.1007/978-3-030-77644-2_111-1

Mishra, P., Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record, 108*(6), 1017–1054. https://doi.org/10.1111/j.1467-9620.2006.00684.x

Mishra, P., Warr, M., & Islam, R. (2023). TPACK in the age of ChatGPT and generative AI. *Journal of Digital Learning in Teacher Education, 39*(4), 235–251. https://doi.org/10.1080/21532974.2023.2247480

Petrillo, J., Sawant, R., Elliott, E., Cleanthous, S., Rogers, R., Cano, S., Johannesen, J., et al. (2024). Rasch measurement theory (RMT) analyses of the Huntington's disease everyday functioning (Hi-DEF) to evaluate item fit and performance. *Journal of Huntington's Disease, 13*(3), 385–397. https://doi.org/10.3233/JHD-230080

Reise, S. P., Hubbard, A. S., Wong, E. F., Schalet, B. D., Haviland, M. G., & Kimerling, R. (2023). Response category functioning on the health care engagement measure using the nominal response model. *Assessment, 30*(2), 375–389. https://doi.org/10.1177/10731911211030164

Saubern, R. (2025). Measuring TPACK learning: A modern measurement approach for TPACK. In M. C. Herring, M. J. Koehler, & P. Mishra (Eds.), *Handbook of Technological Pedagogical Content Knowledge (TPACK) for Educators* (pp. 343–372). Routledge.

Saubern, R., & Urbach, D. (2024). Using item response theory to describe quantitative measures of TPACK. In *Research Highlights in Technology and Teacher Education 2024* (pp. 38–45). Association for the Advancement of Computing in Education.

Schmid, R., Brianza, E., & Petko, D. (2021). Developing a short assessment instrument for technological pedagogical content knowledge (TPACK). *Educational Assessment, Evaluation and Accountability*, *33*(3), 1–27. https://doi.org/10.1016/j.compedu.2020.103967

Tseng, J. J., Chai, C. S., Tan, L., & Park, M. (2022). A critical review of research on technological pedagogical and content knowledge (TPACK) in language teaching.

Computer Assisted Language Learning, 35(4), 948–971. https://doi.org/10.1080/09588221.2020.1868531

Valtonen, T., Sointu, E., Kukkonen, J., Kontkanen, S., Lambert, M. C., & Mäkitalo-Siegl, K. (2023). Examining pre-service teachers' technological pedagogical content knowledge through structural modeling approaches. *Education and Information Technologies*, *28*(4), 4301–4320. https://doi.org/10.1111/jcal.12353

Yeh, Y. F., Chan, K. K. H., & Hsu, Y. S. (2021). Toward a framework that connects individual TPACK and collective TPACK: A systematic review of TPACK studies investigating teacher collaborative discourse in the learning by design process. *Computers & Education, 171*, 104238. https://doi.org/10.1016/j.compedu.2021.104238

Yeşildağ Hasançebi, F., & Günel, M. (2013). The effect of multiple representations on student understanding of physics concepts. *Eurasian Journal of Physics and Chemistry Education*, *5*(1), 5–20. https://doi.org/10.18502/kss.v9i13.16011

Zhao, X., & Wang, F. (2024). A quantitative model of technological pedagogical content knowledge (TPACK) based on graphical calculation. *Discover Education*, *3*, Article 265. https://doi.org/10.18502/kss.v9i13.16011

Zhao, Y., & Wang, J. (2024). Measuring teachers' technological pedagogical content knowledge: Advances, challenges, and future directions. *Education and Information Technologies, 29*(3), 3415–3438.