Indonesian Journal of Educational Development (IJED)
Volume 7, Issue 1, 2026, pp. 410-423
ISSN: 2722-1059 (Online); ISSN: 2722-3671 (Print)
DOI: https://doi.org/10.59672/ijed.v7i1.6354

Measuring TPACK in context: Validating a cross-contextually fair instrument for Indonesian teachers

Uwes Anis Chaeruman*)[1], Suwarno[2], Santi Maudiarti[3], I Gde Wawan Sudatha[4]
[1]Universitas Negeri Jakarta, Jakarta, Indonesia; uwes@unj.ac.id
[2]Universitas Bina Nusantara, Jakarta, Indonesia; suwarno2@binus.edu
[3]Institut Parawisata Trisakti, Jakarta, Indonesia: santimaudiari@gmail.com
[4]Universitas Pendidikan Ganesha, Bali, Indonesia: igdewawans@undiksha.ac.id
*)Corresponding author: Uwes Anis Chaeruman; E-mail addresses: author1@email.com

Article Info

Article history:
Received April 24, 2026
Revised April 30, 2026
Accepted May 12, 2026
Available online May 20, 2026

Keywords: Confirmatory analysis, Differential item functioning, In-service teachers, Pshycometric validation, Technological pedagogical content knowledge

Abstract. As technology integration has become increasingly central to contemporary education, the need for valid and contextually fair instruments for assessing teachers' Technological Pedagogical Content Knowledge (TPACK) has become more urgent. This study aimed to develop and comprehensively validate a TPACK self-assessment instrument for Indonesian in-service teachers. A total of 1,297 teachers selected through purposive and convenience sampling from diverse provinces, school levels, and geographic locations participated in the study. The 48-item instrument, distributed across seven TPACK dimensions (TK, CK, PK, PCK, TCK, TPK, and TPACK), employed a 4-point Likert scale. Psychometric evaluation followed a multistage approach: (a) construct validity via corrected item-total correlations, in which all items met the $r \geq .30$ threshold (range: .57–.79); (b) internal consistency reliability using Cronbach's alpha, with all dimensions exceeding .70 ($\alpha = .78$–.97); (c) exploratory factor analysis (EFA), which supported the theoretical factor structure (KMO = .968; Bartlett's $\chi^2 = 54{,}279.51$, $p < .001$); (d) confirmatory factor analysis (CFA), which yielded marginal fit indices (CFI = .707; RMSEA = .107), suggesting the need for model refinement; and (e) differential item functioning (DIF) analysis, which revealed minimal gender-based bias but notable location-based DIF. Overall, the instrument demonstrated strong psychometric properties for measuring teachers' TPACK in the Indonesian context. The findings highlight the importance of refining TPACK instruments to ensure greater measurement fairness across diverse educational settings and support future teacher professional development initiatives.

Introduction

The integration of technology into classroom instruction has become a central concern in contemporary education, particularly as teachers are increasingly expected to design meaningful technology-enhanced learning experiences. The Technological Pedagogical Content Knowledge (TPACK) framework, initially conceptualized by Mishra & Koehler (2006); Agustika & Diputra (2025) as an extension of Shulman (1986); Gulo et al. (2026) Pedagogical Content Knowledge (PCK), provides a theoretically grounded model for understanding the complex interplay among three fundamental knowledge domains: technology (TK), pedagogy (PK), and content (CK). Since its formulation, TPACK has generated substantial scholarly attention, with more than 1,200 empirical studies published in Scopus-indexed journals between 2006 and 2024 (Schmid et al., 2024). The framework suggests that effective technology-enhanced teaching requires not only mastery of individual knowledge domains, but also an integrated understanding of their intersections, namely TCK, TPK, PCK, and TPACK itself.

The TPACK framework has undergone substantial theoretical refinement over the past two decades. Recent contributions, including the contextualized TPACK model proposed by Petko et al. (2025), have explicitly incorporated a contextual knowledge dimension, acknowledging that teachers' technological-pedagogical decisions are shaped by the institutional, cultural, and infrastructural settings in which they work. Schmid et al. (2024) and Oroh et al. (2025) further argued that TPACK should be conceptualized as a transformative rather than merely additive construct. In this view, TPACK represents an integrated form of professional knowledge in which the interaction among technology, pedagogy, and content cannot be fully understood in terms of separate components. This theoretical development has important implications for TPACK measurement, as instruments based solely on additive assumptions may not adequately capture the integrative nature emphasized in contemporary TPACK research. Consequently, psychometric approaches capable of accommodating both dimension-specific and general factor structures are increasingly needed.

Despite the proliferation of TPACK research, measuring this construct remains a persistent methodological challenge. Schmid et al. (2024), in their comprehensive review, observed that self-assessment instruments continue to dominate TPACK research, although many lack sufficient psychometric rigor. A substantial proportion of existing instruments relies primarily on classical test theory (CTT) indicators, such as Cronbach's alpha, without incorporating more robust analytical approaches, including factor analysis, item response theory (IRT), or differential item functioning (DIF) analysis (Chai et al., 2023). As a result, concerns remain regarding the interpretive validity, structural stability, and cross-contextual fairness of TPACK measurement instruments. These limitations are particularly important in diverse educational settings, where contextual and demographic differences may influence how teachers interpret and respond to TPACK items.

The Indonesian educational context presents unique measurement challenges, further underscoring the need for a rigorously validated TPACK instrument. Indonesia's education system includes more than 3.3 million teachers distributed across a geographically diverse archipelago, with substantial disparities in technological infrastructure between urban and rural areas (Kemendikbudristek, 2023). Recent national surveys indicate that only approximately 35% of teachers in rural areas report confidence in using digital technologies for teaching, a figure considerably lower than that of teachers in urban settings (Astari & Yulianto, 2025). These disparities reflect not only unequal access to technological infrastructure but also differences in professional development opportunities and exposure to pedagogical innovation. Chang et al. (2025) further demonstrated that Indonesian teachers' TPACK trajectories vary across professional development pathways and institutional contexts. Consequently, TPACK instruments intended for use in the Indonesian context must accommodate such heterogeneity while maintaining measurement fairness across demographic and contextual subgroups.

Prior studies on Indonesian teachers' TPACK, including those by Handayani et al. (2024); Sofyan et al. (2023), have provided important initial validation evidence. However, these studies generally relied on relatively small sample sizes and limited psychometric analyses, particularly by omitting differential item functioning (DIF) analyses, which are essential for evaluating instrument fairness across diverse populations. Without evidence of measurement equivalence across geographic, demographic, and professional subgroups, comparisons of TPACK scores may reflect not only actual competency differences, but also construct-irrelevant variance arising from unequal interpretation of instrument items. DIF analysis, which identifies items that function differently across subgroups after controlling for overall ability, has increasingly been recognized as an important methodological approach for ensuring fairness in educational measurement (Effatpanah et al., 2025; Opesemowo, 2025). Nevertheless, DIF analysis remains rarely applied in TPACK validation studies, particularly within the Indonesian educational context. This methodological gap highlights the need for more comprehensive psychometric validation approaches capable of supporting fairer and more contextually appropriate TPACK measurement.

The present study addresses these gaps by developing and comprehensively validating a TPACK assessment instrument using a large-scale sample of 1,297 Indonesian in-service teachers. Unlike many previous TPACK validation studies conducted in the Indonesian context, this study integrates multiple complementary psychometric approaches, including corrected item-total correlations, internal consistency reliability, exploratory factor analysis (EFA), confirmatory factor analysis (CFA), classical item analysis, and differential item functioning (DIF) evaluation, to construct a more comprehensive validity argument. This multilayered framework responds to the growing consensus in educational measurement that no single analytical approach is sufficient to establish construct validity (American Educational Research Association et al., 2014). In particular, the inclusion of DIF analysis enables the study to examine the fairness of TPACK measurement across demographic and contextual subgroups, an aspect that remains underexplored in prior Indonesian TPACK research. Specifically, this study pursues three objectives: (1) to examine the construct validity and dimensional structure of the TPACK instrument, (2) to establish reliability evidence across multiple indicators, and (3) to evaluate the measurement fairness of the instrument across teacher gender and geographic location.

Method

Research Design

This study employed a quantitative cross-sectional survey design to develop and validate a self-assessment instrument for measuring teachers' Technological Pedagogical Content Knowledge (TPACK) in the Indonesian educational context. The study focused on establishing evidence of construct validity, reliability, dimensional structure, and measurement fairness using multiple complementary psychometric approaches. The instrument development and validation process involved several stages, including item development, instrument administration, and statistical evaluation using classical and contemporary psychometric techniques.

Instrument Development

The instrument was developed based on the Technological Pedagogical Content Knowledge (TPACK) framework proposed by Mishra & Koehler (2006) and subsequent developments in TPACK research. Item development was guided by the conceptual dimensions of TPACK, including technological knowledge (TK), pedagogical knowledge (PK), content knowledge (CK), pedagogical content knowledge (PCK), technological content knowledge (TCK), technological pedagogical knowledge (TPK), and technological pedagogical content knowledge (TPACK). Several items were adapted and contextualized to reflect the characteristics of the Indonesian educational setting, particularly variations in technological access, instructional practices, and digital learning environments.

The initial instrument consisted of 48 self-assessment items measured using a four-point Likert scale ranging from strongly disagree to strongly agree. The items were designed to capture teachers' perceptions of their knowledge and ability to integrate technology, pedagogy, and content in instructional practice. Prior to large-scale administration, the instrument underwent refinement in terms of wording clarity, contextual appropriateness, and conceptual alignment with the TPACK framework. The finalized instrument was subsequently administered to Indonesian in-service teachers for psychometric validation.

Participants

A total of 1,297 in-service teachers from various provinces across Indonesia participated in this study. The sample was drawn through purposive and convenience sampling via an online survey platform distributed through provincial education offices and teacher professional networks. The sample's geographic diversity, spanning villages, districts, urban centers, and remote/border areas, was deliberate, ensuring the analytical framework could detect potential measurement non-invariance across the full spectrum of Indonesian teaching contexts. Table 1 presents the demographic characteristics of the participants.

Table 1. Demographic Characteristics of Participants (N = 1,297)

Characteristic	Category	n	%
Gender	Male	453	34.9
	Female	843	65.0
School level	Primary or equivalent	500	38.6
	Junior High or equivalent	300	23.1
	Senior High or equivalent	390	30.1
	Preschool (kindergarten)	82	6.3
	Other	25	1.9
School type	Public	898	69.2
	Private	396	30.5
Geographic location	Village	522	40.2
	District	284	21.9
	Province/City	396	30.5
	Remote/border area	72	5.6
Teaching experience	1–5 years	156	12.0
	6–10 years	472	36.4
	11–15 years	495	38.2
	> 15 years	172	13.3

Note. Exact frequencies are computed from single-selection responses only. One respondent selected both gender options; 23 respondents selected multiple locations; 6 selected multiple school levels; 3 selected multiple school types; 2 selected multiple experience ranges.

The sample size of 1,297 satisfies conventional thresholds for the analyses employed. For EFA, the subject-to-item ratio of 27:1 exceeds the recommended minimum of 10:1 (Hair et al., 2019). For CFA, the sample comfortably surpasses the minimum of 200 recommended by Kline (2016), and for DIF analysis, adequate power is ensured with approximately 400 respondents per group (Zieky, 1993).

Instrument

The TPACK survey consists of 48 self-report items distributed across six dimensions: TK (10 items), CK (3 items), PK (10 items), TCK (8 items), TPK (9 items), and TPACK (8 items). Items are rated on a 4-point Likert scale (1 = Strongly Disagree to 4 = Strongly Agree). The instrument also includes 7 demographic questions and 9 supplementary percentage-based items assessing self-reported frequency of technology-pedagogy-content integration practices, which served as external criteria for concurrent validity assessment. The instrument was developed through a multistage process beginning with item generation grounded in the original Schmidt et al. (2009) TPACK survey and subsequent revisions (Schmid et al., 2024; Sofyan et al., 2023). Items were adapted to the Indonesian cultural and educational context through translation–back-translation procedures and were reviewed by three subject-matter experts in educational technology and instructional design for content validity.

Data Analysis Procedure

Data analysis proceeded through six sequential stages to establish evidence of construct validity, reliability, dimensional structure, and measurement fairness. First, construct validity was evaluated using corrected item-total correlations within each dimension, with items below the $r = .30$ threshold considered for removal (Nunnally & Bernstein, 1994). Second, criterion validity was examined through Pearson correlations between dimension scores and the external criterion measure. Third, reliability was estimated using Cronbach's alpha and Spearman-Brown split-half coefficients for each dimension. Fourth, exploratory factor analysis (EFA) was conducted using principal axis factoring with Varimax rotation to examine the instrument's underlying dimensional structure, preceded by Kaiser-Meyer-Olkin (KMO) and Bartlett's test of sphericity. Fifth, confirmatory factor analysis (CFA) using maximum likelihood estimation was performed to evaluate the extent to which the hypothesized factor structure was supported by the empirical data, based on conventional fit indices including CFI, TLI, RMSEA, and GFI. Finally, differential item functioning (DIF) analysis was conducted using ANOVA-based procedures, with total scores as covariates, to identify items that function differently across demographic subgroups. Items were classified as exhibiting DIF when both statistical significance ($p < .05$) and practical significance (partial $\eta^2 > .01$) criteria were satisfied (Finch & French, 2023).

Results and Discussion

Descriptive Statistics

Table 2 presents the descriptive statistics for each TPACK dimension. Mean scores ranged from 2.795 (TPACK) to 3.356 (CK) on the 4-point scale. The relatively lower means on the integrative dimensions (TCK, TPK, TPACK) compared to the foundational domains (TK, CK, PK) suggest that teachers are more confident in their isolated knowledge areas than in their ability to synthesize these domains in instructional practice. This pattern is consistent with Petko et al. (2025), who argue that the integrative facets of TPACK constitute the most cognitively demanding aspects of teacher professional knowledge, requiring not only mastery of individual components but also the orchestration of those components within authentic pedagogical contexts.

Table 2. Descriptive Statistics by TPACK Dimension (N = 1,297)

Dimension	Items	M	SD
TK	10	2.996	0.431
CK	3	3.356	0.424
PK	10	3.193	0.393
TCK	8	2.837	0.493
TPK	9	2.899	0.470
TPACK	8	2.795	0.502

Note. Scores based on a 4-point Likert scale (1–4).

Construct Validity: Corrected Item-Total Correlations

Table 3 presents the corrected item-total correlation coefficients for all 48 items. Every item achieved a corrected item-total correlation exceeding the .30 minimum threshold, with values ranging from .568 (TK.6) to .794 (PK.6). The consistently high correlations indicate strong internal convergence within each dimension, confirming that individual items adequately contribute to their respective dimension's total score.

Table 3. Corrected Item-Total Correlations by Dimension

Item	r_{it}	Item	r_{it}	Item	r_{it}
TK.1	.605	PK.4	.774	TPK.2	.679
TK.2	.628	PK.5	.683	TPK.3	.706

Item	r_{it}	Item	r_{it}	Item	r_{it}
TK.3	.677	PK.6	.794	TPK.4	.703
TK.4	.614	PK.7	.784	TPK.5	.705
TK.5	.635	PK.8	.763	TPK.6	.666
TK.6	.568	PK.9	.641	TPK.7	.697
TK.7	.659	PK.10	.720	TPK.8	.744
TK.8	.655	TCK.1	.614	TPK.9	.722
TK.9	.650	TCK.2	.628	TPACK.1	.653
TK.10	.665	TCK.3	.666	TPACK.2	.691
CK.1	.593	TCK.4	.761	TPACK.3	.759
CK.2	.680	TCK.5	.766	TPACK.4	.782
CK.3	.597	TCK.6	.713	TPACK.5	.732
PK.1	.707	TCK.7	.721	TPACK.6	.772
PK.2	.717	TCK.8	.590	TPACK.7	.654
PK.3	.748	TPK.1	.681	TPACK.8	.656

Note. All correlations significant at $p < .001$. r_{it} = corrected item-total correlation.

All corrected item-total correlations exceeded the recommended .30 threshold, indicating that each item contributed adequately to its intended dimension. The observed coefficients (.568–.794) were comparable to, and in several cases exceeded, those reported in previous TPACK validation studies. These findings provide strong evidence of internal convergence and support the instrument's construct validity across all dimensions. The EFA results further reinforced the instrument's theoretical structure, with the extracted factors aligning closely with the proposed TPACK framework.

Criterion Validity (Concurrent)

Concurrent validity was assessed by correlating the mean scores for each dimension with an external criterion comprising 9 self-report items measuring teachers' habitual integration of content, pedagogy, and technology in classroom practice. As shown in Table 4, all six dimensions demonstrated statistically significant positive correlations with the criterion measure. The integrative dimensions (TCK, TPK, TPACK) exhibited stronger criterion correlations ($r = .62$–$.65$) than the foundational domains (CK: $r = .31$; PK: $r = .36$), indicating that the composite dimensions capture practical integration behaviors more directly.

Table 4. Concurrent Validity: Correlations Between Dimension Scores and External Criterion

Dimension	r	p	Interpretation
TK	.524	< .001	Strong
CK	.309	< .001	Moderate
PK	.363	< .001	Moderate
TCK	.630	< .001	Strong
TPK	.620	< .001	Strong
TPACK	.647	< .001	Strong

The corrected item-total correlations uniformly exceeded the .30 threshold, with values ranging from .568 to .794. These coefficients are comparable to or exceed those reported in prior TPACK validation studies. For instance, Sofyan et al. (2023) reported item-total correlations ranging from .42 to .76 during the validation of a TPACK instrument for elementary school teachers in Indonesia, while Schmid et al. (2024) obtained values between .51 and .80 during the development of the TPACK scale for pre-service teachers. The EFA results further corroborated the six-dimensional structure, with the Kaiser criterion yielding exactly six factors, directly mirroring the theoretical TPACK framework.

The differential pattern of criterion correlations across dimensions carries important substantive implications. Foundational dimensions (TK, CK, PK) demonstrated weaker relationships with the criterion measure, whereas the integrative dimensions (TCK, TPK, and TPACK) showed substantially stronger correlations. This pattern supports the theoretical proposition that authentic technology integration practices are more closely associated with integrated forms of professional knowledge than with isolated competencies (Chang et al., 2025; Koehler et al., 2013). These findings suggest that integrative TPACK dimensions may provide more meaningful indicators of teachers' actual technology integration practices than standalone knowledge domains.

The high intercorrelations among the TCK, TPK, and TPACK dimensions warrant substantive interpretation rather than being viewed merely as statistical artifacts. Koehler et al. (2013) theorized that TPACK represents a form of knowledge that transcends the simple additive combination of its constituent domains. The empirical difficulty in separating these integrative dimensions supports a transformative rather than additive view of the TPACK framework, a perspective increasingly endorsed in recent literature (Petko et al., 2025; Schmid et al., 2024). These findings suggest that the conceptual boundaries among TCK, TPK, and TPACK may be inherently permeable, particularly when assessed through self-report measures.

Reliability Analysis

Two complementary reliability estimates were computed for each dimension: Cronbach's alpha (α) for internal consistency and Spearman-Brown split-half reliability coefficient (r_SB). Table 5 summarizes these results. All dimensions exceeded the .70 threshold considered acceptable for research instruments (Thomas, 2022). Four dimensions, PK, TPK, TPACK, and the total instrument, achieved coefficients above .90, classified as "excellent" reliability. The CK dimension yielded the lowest alpha (.784), which is expected given its small number of items (k = 3), as alpha is positively influenced by item count.

Table 5. Reliability Coefficients by Dimension

Dimension	k	Cronbach's α	Spearman-Brown r_s[b]	Classification
TK	10	.892	.922	Good
CK	3	.784	.810	Acceptable
PK	10	.932	.952	Excellent
TCK	8	.898	.929	Good
TPK	9	.912	.935	Excellent
TPACK	8	.911	.945	Excellent
Total	48	.972	—	Excellent

Note. k = number of items. Classification follows DeVellis & Thorpe (2022): α ≥ .90 Excellent; .80–.89 Good; .70–.79 Acceptable.

Reliability evidence across both indicators consistently demonstrated adequate to excellent internal consistency for all dimensions. The overall instrument alpha of .972 indicates a very high level of internal consistency and compares favorably with previously reported TPACK instruments. For example, Schmidt et al. (2009) reported alpha coefficients ranging from .75 to .92, while Karataş & Ataç (2025) obtained values between .84 and .93 in their AI-TPACK validation study. The relatively low reliability coefficient for the CK dimension should be interpreted with caution, as internal consistency estimates are strongly influenced by the item count. The corresponding Spearman-Brown coefficient (.810) further supports the adequacy of the CK dimension within the overall instrument structure. Future refinement of the instrument may consider expanding the CK dimension to improve balance across dimensions and potentially enhance reliability and stability.

Exploratory Factor Analysis (EFA)
Prior to factor extraction, the adequacy of the correlation matrix for factor analysis was evaluated. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy was .968, classified as "marvelous" according to Kaiser's (1974) taxonomy, substantially exceeding the .60 minimum. Bartlett's test of sphericity was statistically significant ($\chi^2 = 54{,}279.51$, $p < .001$), confirming that the correlation matrix was not an identity matrix and that factor analysis was warranted.

Table 6. EFA Adequacy Tests

Test	Value	Criterion	Conclusion
KMO	.968	> .60	Marvelous (Highly Adequate)
Bartlett's χ^2	54.279.51	$p < .05$	Significant ($p < .001$)
Factors (eigenvalue > 1)	6	Kaiser Criterion	Consistent with theory

Application of the Kaiser criterion (eigenvalue > 1) yielded six factors that align closely with the theoretical TPACK structure. The first eigenvalue was notably dominant (20.83), accounting for approximately 43.4% of the total variance, followed by a substantial decline to the second factor (eigenvalue = 5.30). This scree pattern suggests the presence of a strong general factor underlying the TPACK construct alongside dimension-specific factors. The extracted factors corresponded to the theoretical TPACK domains, although cross-loading patterns were observed among items from the integrative dimensions (TCK, TPK, and TPACK), a finding consistent with prior TPACK research (Archambault & Barnett, 2010; Schmid et al., 2024).

The dominant first eigenvalue (43.4% of variance) is theoretically meaningful within the contemporary TPACK literature. Petko et al. (2025) and Schmid et al. (2024) have argued that the empirical inseparability of TPACK dimensions reflects not measurement failure but the construct's intrinsic transformational nature; teachers who develop competence in one dimension tend to cultivate competence across all dimensions concurrently. The present scree pattern is therefore consistent with a hierarchical conception of TPACK in which a general factor coexists with dimension-specific factors, supporting the case for bifactor or higher-order modeling in future analyses.

Confirmatory Factor Analysis (CFA)
A six-factor CFA model was specified, with each observed item loading exclusively on its hypothesized latent factor. The model was estimated using maximum likelihood (ML) estimation. All factor loadings were statistically significant ($p < .001$), with standardized estimates presented in Table 7. Table 8 reports the goodness-of-fit indices.

Table 7. CFA Standardized Factor Loadings (Selected Items)

Factor	Item	Estimate	SE	z	p
TK	TK.1 (reference)	1.000	—	—	—
TK	TK.3	1.172	0.057	20.39	< .001
TK	TK.8	1.626	0.075	21.63	< .001
TK	TK.10	1.639	0.076	21.62	< .001
CK	CK.1 (reference)	1.000	—	—	—
CK	CK.2	1.089	0.047	23.26	< .001
PK	PK.6	1.088	0.036	30.56	< .001
PK	PK.7	1.078	0.036	30.33	< .001
TCK	TCK.5	1.714	0.067	25.57	< .001
TPK	TPK.5	1.712	0.066	26.03	< .001
TPACK	TPACK.4	1.478	0.052	28.28	< .001
TPACK	TPACK.6	1.445	0.053	27.29	< .001

Note. Reference indicators fixed at 1.000. SE = standard error.

Table 8. CFA Goodness-of-Fit Indices

Index	Obtained	Acceptable Threshold	Evaluation
χ^2 (df = 1065)	16,841.18	p > .05 (sample-sensitive)	Significant
CFI	.707	≥ .90 (good); ≥ .95 (excellent)	Below threshold
TLI	.690	≥ .90	Below threshold
GFI	.694	≥ .90	Below threshold
RMSEA	.107	≤ .08 (good); ≤ .05 (excellent)	Above threshold
AIC	196.03	Lower = better (comparative)	—

The CFA results indicate that the hypothesized six-factor model did not fully achieve conventional goodness-of-fit thresholds (Hu & Bentler, 1999). The indices obtained (CFI = .707; RMSEA = .107) suggest that the strict separation of the six TPACK dimensions may not fully reflect the empirical covariance structure of the data. However, this pattern is not uncommon in TPACK validation research, particularly when the framework is operationalized as multiple closely related dimensions through self-report measures. Similar challenges in achieving optimal fit for multidimensional TPACK models have been reported in prior studies (Archambault & Barnett, 2010; Schmid et al., 2024). The significant chi-square statistic (χ^2 = 16,841.18) should also be interpreted cautiously, given the sensitivity of chi-square tests to large sample sizes.

The marginal CFA fit indices (CFI = .707; RMSEA = .107) indicate that the standard six-factor specification does not optimally represent the data's covariance structure. This finding, while initially concerning, is consistent with a substantial body of TPACK CFA literature. Archambault & Barnett (2010) reported similar difficulties in confirming the seven-factor TPACK model, and Chai et al. (2013) noted that alternative specifications, such as higher-order or bifactor models, often yield superior fit. The suboptimal fit in the present study may reflect several interrelated factors: (a) the ordinal nature of Likert data, which violates the continuous-variable assumption of ML estimation; (b) the high dimensionality of the model relative to the number of items per factor; and (c) the inherent conceptual overlap among integrative dimensions.

These results argue for exploring alternative model specifications in future research. A second-order model, in which a general TPACK factor explains the covariances among the six first-order factors, may provide a more parsimonious and better-fitting representation. Alternatively, a bifactor model that partitions variance into a general factor and dimension-specific factors could help disentangle the common and unique variance components, addressing both the strong general factor evident in the scree plot and the dimension-specific reliability demonstrated by the alpha coefficients. Such modeling approaches are increasingly recommended for multidimensional psychological constructs that exhibit substantial inter-dimensional correlations, as is characteristic of integrative knowledge frameworks such as TPACK.

Classical Item Analysis (IRT Proxy)

Classical item analysis was conducted using the upper-lower 27% group method as a practical approximation of IRT parameters. The difficulty index (p) represents the ratio of the item to the maximum possible score, while the discrimination index (D) measures the difference in endorsement rates between high- and low-scoring respondent groups, normalized by the maximum score range.

Table 9. Integrated Psychometric Profile: Selected Items

Item	r_{it}	p	Category	D	Category
TK.7	.659	.632	Moderate	.388	Good
TK.10	.665	.686	Moderate	.413	Excellent
CK.1	.593	.880	Easy	.332	Good
PK.4	.774	.792	Easy	.283	Fair

Item	r_{it}	p	Category	D	Category
PK.6	.794	.805	Easy	.293	Fair
TCK.4	.761	.640	Moderate	.393	Good
TCK.5	.766	.641	Moderate	.403	Excellent
TPK.5	.705	.637	Moderate	.405	Excellent
TPACK.4	.782	.635	Moderate	.405	Excellent
TPACK.6	.772	.669	Moderate	.414	Excellent

Note. r_{it} = corrected item-total correlation; p = difficulty index; D = discrimination index (27% upper-lower method). Difficulty: Easy ≥ .76; Moderate = .25–.75. Discrimination: Excellent ≥ .40; Good = .30–.39; Fair = .20–.29.

Across all 48 items, difficulty indices ranged from .598 (TPK.6) to .880 (CK.1), indicating that most items were perceived as relatively easy by respondents, a pattern consistent with the tendency toward positive self-assessment in Likert-based instruments (Jordan & Troth, 2020; Paap et al., 2023; Widana et al., 2021). Items from the integrative dimensions (TCK, TPK, TPACK) tended to be moderately difficult, reflecting the greater cognitive complexity of synthesizing multiple knowledge domains. All items achieved discrimination indices at or above .20, with four items exceeding the .40 threshold for excellent discrimination: TK.10 (D = .413), TCK.5 (D = .403), TPK.5 (D = .405), TPACK.4 (D = .405), and TPACK.6 (D = .414).

The pattern of moderate-to-easy difficulty indices, coupled with adequate-to-excellent discrimination indices, indicates that the instrument's items serve as effective markers of TPACK self-perception across the ability range. The clustering of higher discrimination values among integrative items (TCK.5, TPK.5, TPACK.4, TPACK.6) suggests that responses to integrative items are more sensitive to genuine differences in respondents' overall TPACK level than responses to foundational items, where ceiling effects and social desirability bias may compress score variance (Paap et al., 2023). Integrative items, therefore, appear to carry the most diagnostic information and should be preserved or expanded in future revisions.

Differential Item Functioning (DIF)

DIF analysis was performed across two grouping variables: teacher gender and geographic location. Total scores served as the matching criterion to control ability differences between groups. Items were classified as exhibiting DIF when both statistical significance ($p < .05$) and practical significance (partial $\eta^2 > .01$) were met simultaneously.

Gender-Based DIF

Only two of 48 items (4.2%) exhibited statistically and practically significant DIF across gender groups. Both items belonged to the TK dimension: TK.2 ($\eta^2 = .012$) and TK.5 ($\eta^2 = .013$). The effect sizes were small, suggesting that the magnitude of gender-based measurement bias is negligible and does not compromise the instrument's overall fairness for cross-gender comparisons.

Location-Based DIF

A more pronounced DIF pattern emerged across geographic location groups (village vs. district). Sixteen of 48 items (33.3%) met both significance criteria. This finding indicates that a substantial proportion of items functioned differently across geographic contexts despite comparable overall TPACK levels. Table 10 presents the items exhibiting location-based DIF.

Table 10. Items Exhibiting Location-Based DIF (Village vs. District)

Item	Partial η^2	Item	Partial η^2
TK.5	.015	PK.8	.026
TK.10	.015	PK.10	.015

Item	Partial η^2	Item	Partial η^2
PK.1	.014	TCK.2	.016
PK.3	.018	TCK.5	.025
PK.4	.025	TCK.7	.019
PK.5	.033	TPK.2	.021
PK.6	.020	TPACK.6	.020
PK.7	.017	TPACK.8	.018

Note. DIF criterion: $p < .05$ and partial $\eta^2 > .01$. PK.5 exhibited the largest effect size ($\eta^2 = .033$).

The DIF analysis revealed a clear contrast between gender-based and location-based measurement invariance. The near-absence of gender DIF (2 items; 4.2%) is encouraging and consistent with findings from Valtonen et al. (2019), who reported minimal gender-related DIF in their Finnish TPACK instrument. These results suggest that the present instrument functions relatively equivalently across male and female teachers, supporting valid cross-gender comparisons.

In contrast, the substantial location-based DIF (16 items; 33.3%) represents one of the most important findings of this study. The disproportionate concentration of DIF items within the PK dimension suggests that contextual factors may systematically influence how pedagogical knowledge items are interpreted and endorsed. Teachers in village settings may have varying levels of exposure to formal pedagogical training frameworks, instructional technologies, or standardized teaching practices, which may shape their interpretation of the pedagogical terminology and classroom practices referenced in the instrument. At the same time, the observed DIF pattern may also reflect genuine contextual variation in pedagogical practice rather than purely measurement bias. This interpretation is particularly relevant in geographically diverse educational systems such as Indonesia, where instructional realities may differ substantially across regions.

The pronounced location-based DIF aligns with broader evidence documenting disparities in teacher professional development and educational resources between rural and urban regions in Indonesia (Astari & Yulianto, 2025). Contemporary DIF literature emphasizes that differential item functioning should not be automatically interpreted as item deficiency but rather as an opportunity to examine how contextual conditions influence assessment responses (Chen et al., 2025; Effatpanah et al., 2025). The present findings, therefore, highlight the importance of conducting qualitative follow-up investigations to better understand how teachers from different geographic contexts interpret pedagogical terminology and instructional scenarios embedded within TPACK instruments.

These findings carry important implications for educational policy and large-scale teacher assessment. The Indonesian Ministry of Education initiative aims to improve educational quality more equitably across regions (Kemendikbudristek, 2023). If measurement instruments function differently across geographic contexts, policy decisions based on those assessments may inadvertently misrepresent teachers' actual competencies. The present results, therefore, underscore the importance of incorporating DIF analysis as a standard psychometric procedure when instruments are administered across demographically heterogeneous populations.

Beyond the Indonesian context, these findings contribute to the growing international discussion regarding equitable measurement in teacher assessment. Recent studies in educational and language assessment have demonstrated that ignoring DIF can lead to systematically biased interpretations (Chen et al., 2025) and potentially disadvantage already marginalized populations (Opesemowo et al., 2025). The present study provides additional empirical support for integrating DIF analysis into large-scale teacher competency assessment, particularly in educational systems characterized by substantial infrastructural, geographic, and socio-cultural diversity.

Conclusion

The 48-item TPACK instrument developed in this study demonstrated generally strong psychometric properties across multiple validation procedures. All items met construct validity criteria, reliability coefficients ranged from good to excellent across dimensions, and the six-factor structure was supported by the EFA results. Although the CFA findings indicated marginal model fit, the overall pattern of results suggests that the instrument provides a useful and theoretically grounded measure of Indonesian in-service teachers' TPACK. At the same time, the substantial location-based DIF identified in several items indicates that further refinement is needed to improve measurement fairness across geographic contexts. The 48-item TPACK instrument developed in this study demonstrates strong psychometric properties across multiple validation criteria. All items met construct validity thresholds, reliability coefficients ranged from good to excellent across all dimensions, and the six-factor structure was empirically confirmed through EFA. The instrument is suitable for assessing in-service teachers' TPACK in the Indonesian context, with the caveat that location-based DIF in a substantial proportion of items warrants targeted revision before the instrument is used for cross-regional comparative purposes. This study contributes to TPACK measurement scholarship in three principal ways. First, it provides one of the largest psychometric validations of a TPACK self-assessment instrument in the Indonesian context, with a sample size more than three times that of comparable national studies. Second, it demonstrates the feasibility and value of integrating multiple psychometric perspectives, classical test theory, factor analysis, classical item analysis, and DIF, within a single validation study, offering a methodological template for future instrument development efforts. Third, it offers preliminary empirical evidence supporting the transformative interpretation of TPACK by showing that integrative dimensions (TCK, TPK, TPACK) carry stronger criterion validity and discrimination power than foundational dimensions. The findings also carry important implications for teacher assessment and educational policy. The presence of substantial location-based DIF suggests that contextual factors may influence how teachers interpret pedagogical and technology-related items, particularly across geographically diverse educational settings. Consequently, future revisions of the instrument should prioritize refining pedagogical knowledge items that exhibit stronger DIF effects. More broadly, the study highlights the importance of incorporating fairness-oriented psychometric procedures, such as DIF analysis, into large-scale teacher competency assessments. Several limitations should be acknowledged. The use of a cross-sectional self-report design may introduce response biases and does not allow examination of TPACK development over time. In addition, the online survey approach may have underrepresented teachers in remote regions with limited digital access. Future studies are encouraged to explore higher-order or bifactor CFA models, apply item response theory (IRT) approaches, and conduct qualitative investigations into the interpretation of DIF-sensitive items across different educational contexts. The integration of emerging AI-related dimensions into future TPACK instruments may also provide valuable directions for subsequent research.

Bibliography

Agustika, G. N. S., & Diputra, K. S. (2025). Effectiveness of hybrid project-based learning with digital portfolios in enhancing mathematics pedagogical content knowledge. *Indonesian Journal of Educational Development (IJED)*, *6*(2), 293–308. https://doi.org/10.59672/ijed.v6i2.4923

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

Archambault, L. M., & Barnett, J. H. (2010). Revisiting technological pedagogical content knowledge: Exploring the TPACK framework. *Computers & Education*, *55*(4), 1656–1662.

https://doi.org/10.1016/j.compedu.2010.07.009

Astari, & Yulianto, D. (2025). Bridging the digital divide in education: Disparities in Google Classroom utilization and technical challenges among urban and rural teachers. *Journal of Education Technology*, *9*(2). https://doi.org/10.23887/jet.v9i2.92897

Chai, C. S., Koh, J. H. L., & Tsai, C.-C. (2013). A review of technological pedagogical content knowledge. *Educational Technology & Society*, *16*(2), 31–51.

Chai, C. S., Koh, J. H. L., & Tsai, C.-C. (2023). A review of technological pedagogical content knowledge. *Educational Technology & Society*, *26*(2), 31–51.

Chang, C.-F., Annisa, N., & Chen, K.-Z. (2025). Pre-service teacher professional education program (PPG) and Indonesian science teachers' TPACK development: A career-path comparative study. *Education and Information Technologies*, *30*(7), 8689–8711. https://doi.org/10.1007/s10639-024-13160-6

Chen, X., Aryadoust, V., & Zhang, W. (2025). A systematic review of differential item functioning in second language assessment. *Language Testing*, *42*(2), 213–238. https://doi.org/10.1177/02655322241290188

Effatpanah, F., Ravand, H., & Doebler, P. (2025). Differential item functioning analysis of Likert scales: An overview and demonstration of rating scale tree model. *Psychological Reports*. https://doi.org/10.1177/00332941241308806

Finch, W. H., & French, B. F. (2023). Effect sizes for estimating differential item functioning influence at the test level. *Psych*, *5*(1), 133–147. https://doi.org/10.3390/psych5010013

Gulo, R. S., Niswanto, Ismail, & Handayani, N. (2026). Coaching-based academic supervision to enhance teachers' pedagogical competence. *Indonesian Journal of Educational Development (IJED)*, *6*(4), 1294–1305. https://doi.org/10.59672/ijed.v6i4.5683

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Cengage Learning.

Handayani, S., Hussin, M., & Norman, H. (2024). Evaluating teaching readiness using the TPACK model: Factor, reliability and validity analyses for Indonesian economics teacher candidates. *Perspectives of Science and Education*, *68*(2), 679–698. https://doi.org/10.32744/pse.2024.2.41

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Jordan, P. J., & Troth, A. C. (2020). Common method bias in applied settings: The dilemma of researching in organizations. *Australian Journal of Management*, *45*(1), 3–14. https://doi.org/10.1177/0312896219871976

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, *39*(1), 31–36. https://doi.org/10.1007/BF02291575

Karataş, F., & Ataç, B. A. (2025). When TPACK meets artificial intelligence: Analyzing TPACK and AI-TPACK components through structural equation modelling. *Education and Information Technologies*, *30*(7), 8979–9004. https://doi.org/10.1007/s10639-024-13164-2

Kemendikbudristek. (2023). *Buku saku Merdeka Belajar [Merdeka Belajar handbook]*. Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). The Guilford Press.

Koehler, M. J., Mishra, P., & Cain, W. (2013). What is technological pedagogical content knowledge (TPACK)? *Journal of Education*, *193*(3), 13–19. https://doi.org/10.1177/002205741319300303

Megawaty, M., Sulistiyo, U., Wachyuni, S., & Sofyan. (2025). Development of M.E.G.A: A self-assessment expert system to improve English correspondence skills for job-seeking university students. Indonesian Journal of Educational Development (IJED), 6(2), 393–407. https://doi.org/10.59672/ijed.v6i2.5099

Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record*, *108*(6), 1017–1054.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

Opesemowo, O. A. G. (2025). Exploring undue advantage of differential item functioning in high-stakes assessments: Implications on sustainable development goal 4. *Social Sciences & Humanities Open*, *11*, 101257. https://doi.org/10.1016/j.ssaho.2024.101257

Oroh, S. S. A., Arsana, I. K. S., & Dolonseda, H. P. (2025). Self-efficacy as a moderator of the effect of techno-pedagogical competence on economics teachers' readiness. *Indonesian Journal of Educational Development (IJED)*, *6*(3), 1116–1130. https://doi.org/10.59672/ijed.v6i3.4929

Paap, K. R., Anders-Jefferson, R. T., Balakrishnan, N., & Majoubi, J. B. (2023). The many foibles of Likert scales challenge claims that self-report measures of self-control are better than performance-based measures. *Behavior Research Methods*, *56*(2), 908–933. https://doi.org/10.3758/s13428-023-02089-2

Petko, D., Mishra, P., & Koehler, M. J. (2025). TPACK in context: An updated model. *Computers and Education Open*, *8*, 100244. https://doi.org/10.1016/j.caeo.2025.100244

Schmid, M., Brianza, E., Mok, S. Y., & Petko, D. (2024). Running in circles: A systematic review of reviews on technological pedagogical content knowledge (TPACK). *Computers & Education*, *214*, 105024. https://doi.org/10.1016/j.compedu.2024.105024

Schmidt, D. A., Baran, E., Thompson, A. D., Mishra, P., Koehler, M. J., & Shin, T. S. (2009). Technological pedagogical content knowledge (TPACK): The development and validation of an assessment instrument for preservice teachers. *Journal of Research on Technology in Education*, *42*(2), 123–149. https://doi.org/10.1080/15391523.2009.10782544

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15*(2), 4–14. https://doi.org/10.3102/0013189X015002004

Sofyan, S., Habibi, A., Sofwan, M., Yaakob, M. F. M., Alqahtani, T. M., Jamila, A., & Wijaya, T. T. (2023). TPACK–UotI: The validation of an assessment instrument for elementary school teachers. *Humanities and Social Sciences Communications*, *10*(1), 55. https://doi.org/10.1057/s41599-023-01533-0

Thomas, S. (2022). Robert F. DeVellis & Carolyn T. Thorpe (2022). Scale development: Theory and applications (Book Review). *Personnel Psychology*, *75*(1), 243–244. https://doi.org/10.1111/peps.12499

Valtonen, T., Sointu, E., Kukkonen, J., Mäkitalo, K., Hoang, N., Häkkinen, P., Järvelä, S., Näykki, P., Virtanen, A., Pöntinen, S., Kostiainen, E., & Tondeur, J. (2019). Examining pre-service teachers' technological pedagogical content knowledge as evolving knowledge domains: A longitudinal approach. *Journal of Computer Assisted Learning*, *35*(4), 491–502. https://doi.org/10.1111/jcal.12353

Widana, I. W., Sopandi, A. T., Suwardika, I. G. (2021). Development of an authentic assessment model in mathematics learning: A science, technology, engineering, and mathematics (STEM) approach. *Indonesian Research Journal in Education,* *5*(1), 192-209. https://doi.org/10.22437/irje.v5i1.12992

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Lawrence Erlbaum Associates.