Jurnal Ilmu Pendidikan dan Pembelajaran
e-ISSN: 2962-6579
p-ISSN: 2962-3464

Vol 04, No 02, 228-239
https://doi.org/10.58706/jipp
jipp.mep@gmail.com

Pilot Testing the Internal Structure of the Multicultural Competency Scale for Prospective Teachers

Wahyu Widodo [1,*] and Sugiantoro [2]
[1] Program Studi S2 Psikologi, Universitas Negeri Malang, Indonesia
[2] Program Studi S1 Pendidikan IPS, Universitas Negeri Surabaya, Indonesia
[*]Email: wahyu.widodo.fpsi@um.ac.id

Abstract

The assessment of pre-service teachers' multicultural competence is considered fundamental for the realization of inclusive education; however, such assessment requires robust psychometric instruments. In this study, the construct validity of a Multicultural Competence Scale was evaluated using Confirmatory Factor Analysis (CFA). Following rigorous preliminary diagnostics to ensure the suitability of the data, the structural model was subsequently evaluated in terms of overall fit, validity, and reliability. Although a strong structural foundation was demonstrated by the overall measurement model, several localized weaknesses were identified within the awareness dimension. In particular, items A1 and A6 were found to fall below the required factorial thresholds (loadings < 0.30), indicating the presence of contextual and linguistic misalignment that needs to be revised. Ultimately, these findings highlight a key methodological tension, in which a balance must be maintained between comprehensive construct representation and item efficiency. An empirical foundation for improving the instrument's cultural sensitivity and structural integrity prior to large-scale application is thus provided by this study.

Keywords: *Confirmatory Factor Analysis; Multicultural Competency; Pilot Study; Prospective Teachers*

Manuscript History
Received: December 19, 2025 Revised: March 10, 2026 Accepted: April 23, 2026

How to cite:
Widodo, W. and Sugiantoro. (2026). Pilot Testing the Internal Structure of the Multicultural Competency Scale for Prospective Teachers. *Jurnal Ilmu Pendidikan dan Pembelajaran*, **4**(2), 228-239. DOI: https://doi.org/10.58706/jipp.v4n2.p228-239.

INTRODUCTION

Today's classrooms are characterized by the coexistence of diverse cultures, beliefs, and backgrounds, so teachers can no longer be positioned merely as subject experts. The ability to manage this diversity is seen as a key factor in fostering tolerance and harmonious interaction among students. Multicultural perspectives are now expected to be integrated naturally by teachers in all aspects of their roles, whether in delivering subject matter or shaping students' character from early childhood through senior education (Pratama et al., 2023; Pratiwi et al., 2024). As highlighted by Ningtyas (2024), this approach is considered the most effective way to prepare students for real-world social dynamics. Although educational trends have increasingly emphasized student-centered learning, the teacher's role continues to be positioned as the central anchor, through which the tone of a fair and inclusive classroom environment is established.

The capacity to manage student diversity and to foster respectful learning environments is largely determined by the level of multicultural competence that is possessed by teachers. However, a persistent issue has been identified in the literature: even though positive attitudes toward multicultural education are often demonstrated, effective implementation in practice is still frequently not achieved (Fatmawati et al., 2023; Amarullah et al., 2024). This gap is commonly associated with a limited understanding of multicultural strategies and insufficient access to professional resources, both of which hinder the development of genuinely inclusive classrooms (Afifuddin et al., 2025). As a result, the development of robust and psychometrically

sound assessment instruments is considered essential, particularly for pre-service teachers who will soon encounter highly diverse student populations.

With more than 300 ethnic groups, approximately 200 languages, and six major religions, both a strength and a unique challenge are presented by Indonesia's demographic composition. As noted by Mariyono (2024), proper management of diversity will strengthen unity and tolerance; while neglecting to manage it could potentially trigger division. Multicultural education in Indonesia itself has emerged not from formal obligations, but from the internalization of values of harmony that have been deeply rooted in the culture since the era of independence (Khalim and Parut, 2025; Maryono, 2024). Its orientation is clear: to create a peaceful pluralistic environment so that all elements of society have equal rights in pursuing well-being. Achieving this goal requires that multicultural competence be not merely a requirement but a foundational skill that prospective teachers must master, given the anticipated complexity of future student populations (Khalim and Parut, 2025).

Efforts to enhance the competencies of prospective educators heavily depend on the availability of high-quality assessment tools. A good assessment tool will yield precise evaluations while facilitating the development of a focused training curriculum. In the literature, there are several popular instruments such as the MTCS (Spanierman et al., 2011) and the MCS (Erdem, 2020). The MCS is considered most suitable for the context of pre-service teacher education because it maps competencies into three specific areas: attitudes, knowledge, and skills. Drawing on its established validity and reliability, the MCS was adapted into Indonesian by Widodo and Chotimah (2023), and its validation was conducted using the Rasch Measurement Model.

Through the adaptation and validation process, important insights into item quality, one-dimensionality, and fairness at both the item and respondent levels were identified. The accuracy of the measurement instrument was confirmed through Rasch analysis, but its reliance on probability-based methods limited its ability to fully evaluate the construct model against observed data. Therefore, the results of the MCS adaptation and validation still need to be further examined to ensure their appropriateness from a latent variable measurement perspective, particularly by applying Confirmatory Factor Analysis (CFA). Umar and Nisa (2020) regard CFA as the most reliable method for evaluating construct validity in psychology, education, and the social sciences. CFA enables researchers to assess the degree to which items within an instrument accurately measure the intended construct, and is applicable to both unidimensional and multidimensional models. Additionally, CFA facilitates the evaluation of the fit between theoretical models and empirical data through various fit indices, including Chi-Square, RMSEA, CFI, and TLI. In essence, CFA serves as a confirmatory psychometric analysis method, complementing the findings of the previous Rasch analysis.

Given this urgency, this study aimed to test the construct validity of the instrument using CFA. Meanwhile, considering the large number of respondents required for CFA and the complex analysis results, to anticipate procedural weaknesses and errors in adapting items, this study was conducted as a pilot study. By conducting a CFA analysis within a pilot study design, this study bridges the gap between Rasch model-based and structural model-based validation. The present study reports methodological results, and also sets the ground for a broader psychometric analysis of a multicultural competence measure for pre-service teachers based on measurement model approaches. Not only does this study presents findings related to methodology, it also lays the groundwork for a more thorough psychometric exploration of the prospective teacher multicultural competence measure from multiple measurement model perspectives. This base can be used later for the next research and development of multicultural education in Indonesia.

METHOD

A quantitative approach was chosen for this study, relying on Confirmatory Factor Analysis (CFA) to strictly test the construct validity of the 14 adapted items. The goal was to see exactly how well these items mapped onto the target constructs. Before executing the CFA, the dataset was screened through a series of prerequisites: descriptive statistics, univariate and multivariate normality checks, and structural adequacy confirmed through the KMO index and Bartlett's test. With the data validated, the CFA was deployed using a Maximum Likelihood Robust (MLR) estimator. The subsequent steps involved evaluating overall model fit, convergent and discriminant validity, and structural reliability via Composite Reliability (CR) and Average Variance Extracted (AVE). In addition to the statistical procedures, the demographic profile of the participants is detailed in a table to explicitly outline the sample's cultural and gender composition, providing necessary background for the upcoming analysis.

this study using quantitative approach with Confirmatory Factor Analysis (CFA). The CFA applied to rigorously assess the construct validity of the 14 adapted items. The extent to which these items aligned with

the intended constructs was then specifically examined. Prior to conducting the CFA, a series of prerequisite tests were apllied in dataset, including descriptive statistical analysis, assessments of univariate and multivariate normality, and evaluation of structural adequacy using the KMO index and Bartlett's test.

After the data were deemed suitable, CFA was performed using the Maximum Likelihood Robust (MLR) estimator. Subsequently, the overall model fit was evaluated, along with convergent and discriminant validity, as well as structural reliability, which was assessed using Composite Reliability (CR) and Average Variance Extracted (AVE). In addition to these statistical procedures, the demographic characteristics of the participants were presented in tabular form to clearly describe the cultural and gender composition of the sample, thereby providing essential context for the subsequent analysis.

Table 1. Demographic Diversity of Participants

Category	Sub Category	n	Percentage (%)
Gender	Male	25	25%
	Female	77	75%
Ethnicity	Java	41	40%
	Dayak	9	9%
	Madura	9	9%
	Manggarai	8	8%
	Sumba	7	7%
	Bunak	5	5%
	Ende	5	5%
	Timor	4	4%
	Malaka	2	2%
	Tetun	2	2%
	Fehan	1	1%
	Klisok Lor	1	1%
	Kodi	1	1%
	Leleng	1	1%
	Mbaling	1	1%
	Mbojo	1	1%
	Metang	1	1%
	Moor	1	1%
	Nias	1	1%
	Sandiata	1	1%
Education Program	Undergraduate students in teacher education programs (East Java Universities)	102	100%

Although the sample size (N = 102) may be considered relatively small for second-order Confirmatory Factor Analysis, it still meets the minimum acceptable criteria for CFA using the Maximum Likelihood Robust (MLR) estimator, which is considered adequate for models with moderate complexity and factor loadings above recommended thresholds (Hair et al., 2019; Kline, 2016). As shown in Table 1, the respondents involved in this study comprised to 102 participants, The participants consisted of 102 students, with 25% male and 75% female, representing diverse ethnic backgrounds predominantly from Java (40%) along with various other ethnic groups in smaller proportions. The respondents involved in this study came from universities in East Java who were pursuing undergraduate education in educational study programs. Respondents were reached using convenience sampling techniques, considering the ease of implementation.

The Multicultural Competence Scale (MCS) for prospective teachers includes awareness, knowledge, and skills. The scale was adapted to Indonesian through cross-cultural adaptation by Erdem (2020). After expert review and EFA, the scale had 14 items. The second-order CFA showed adequate fit indices with significant factor loading. Measurement invariance testing proved that the instrument could measure multicultural competence without gender bias. The MCS was adapted by Widodo and Chotimah (2023) using Beaton et al. (2000)'s technique through translation, synthesis, back-translation, expert assessment, and testing. Rasch analysis examined respondent quality, item quality, and instrument quality. Seven items were eliminated as misfits, leaving seven final items. The scale met the unidimensionality criteria with 76.2% Raw Variance

Explained and high item reliability. Wright Map analysis showed ideal item-respondent distribution, and rating scale analysis confirmed optimal five-point Likert scale functioning.

Procedure and Data Analysis

The dataset is the same as that used by Widodo and Chotimah (2023). Data analysis assessed construct validity using Confirmatory Factor Analysis (CFA). Descriptive statistical tests, normality tests, and data feasibility using the Kaiser-Meyer-Olkin (KMO) Measure and Bartlett's Test of Sphericity were conducted. A KMO value > 0.60 indicates data appropriateness (Kaiser, 1974). Bartlett's test was used to analyze correlations with a significance value of <0.05 for factor analysis (Hair, 2019). Model evaluation examined goodness-of-fit indices: Chi-square ($\chi 2$), RMSEA, CFI, and TLI (Jackson et al., 2009). Construct validity was evaluated using two approaches. First, convergent validity was based on factor loading values (≥ 0.30) (Indu et al., 2025), AVE (≥ 0.50) (Fornell and Larcker, 1981), and CR (≥ 0.70) (Hair et al., 2019). Second, discriminant validity compared the square root of the AVE with the construct correlation using the HTMT analysis (Henseler et al., 2015). Analyses were conducted using R Programme version 4.5.0 via R Studio 2025.05.0, with statistical packages MVN version 6.1, Psych version 2.5.3, and lavaan version 0.6-19.

RESULTS AND DISCUSSION

As a preliminary step to validate the dataset (Harbison and Simmons, 2024), descriptive statistics were run on the 14 items. Scores ranged from 1.43 to 4.22, highlighting a clear divide: knowledge items K1 (M = 4.17, SD = 1.03) and K3 (M = 4.22, SD = 1.03) scored highest, while awareness items A1 (M = 1.43, SD = 0.95) and A6 (M = 1.44, SD = 0.97) scored lowest. Distribution-wise, the data is moderately acceptable. Most items met normality criteria (skewness: −1 to +1; kurtosis: −2 to +2), though specific asymmetries existed. Notably, A1 and A6 were leptokurtic with high positive skewness, whereas K1 and K3 skewed negatively. Complete item descriptions can be found in Table 2.

Table 2. Descriptive Statistics of Intercultural Competence Scale Items

Code	Item wording	Mean	SD	Skewness	Kurtosis	Note
A1	*Budaya saya menjauhkan saya dari siswa yang memiliki latar belakang budaya yang berbeda.* My culture distances me from students with different cultural backgrounds.	1.43	0.95	2.31	4.71	Negative item; high skewness
A2	*Saya memahami berbagai karakteristik budaya dari masing-masing siswa.* I understand the various cultural characteristics of each student.	3.29	1.13	-0.30	-0.59	-
A3	*Saya merasa insyaf bila saya telah mendiskriminasi siswa yang berlatar belakang budaya yang berbeda.* I feel sorry if I have discriminated against students from different cultural backgrounds.	3.20	1.40	-0.28	-1.25	-
A4	*Saya mampu mengkritisi prasangka dalam diri saya pribadi yang muncul akibat pengalaman interaksi dengan budaya yang beraneka ragam.* I can criticize my own personal prejudices that arise from my experiences interacting with diverse cultures.	3.43	1.17	-0.54	-0.51	-
A5	*Saya mengetahui adanya prasangka yang muncul dalam diri saya pribadi akibat adanya interaksi dengan budaya yang beraneka ragam.* I am aware of the prejudices that arise within me due to my interactions with diverse cultures.	3.36	1.08	-0.37	-0.54	-
A6	*Budaya yang saya miliki membuat saya berperilaku melawan terhadap siswa yang berlatar belakang budaya yang berbeda.*	1.44	0.97	2.27	4.38	Negative item; high skewness

Code	Item wording	Mean	SD	Skewness	Kurtosis	Note
	The culture I have makes me behave rebelliously towards students from different cultural backgrounds.					
S1	*Saya mampu merancang lingkungan belajar yang edukatif yang sesuai dengan latar belakang budaya siswa yang beraneka ragam.* I am able to design an educational learning environment appropriate to students' cultural diversity.	3.77	0.95	-0.47	-0.39	-
S2	*Saya mampu menyusun soal ujian yang sesuai dengan latar belakang budaya siswa yang beraneka ragam.* I am able to develop exam questions appropriate to students' cultural diversity.	3.43	1.14	-0.45	-0.47	-
S3	*Saya mampu menyusun materi pelajaran yang sesuai dengan latar belakang budaya siswa yang beraneka ragam.* I am able to develop learning materials that are appropriate to students' diverse cultural backgrounds.	3.41	1.16	-0.60	-0.39	-
S4	*Saya mampu mengelola pembelajaran secara tepat sesuai dengan latar belakang budaya siswa yang beraneka ragam.* I am able to manage learning according to students' cultural diversity.	3.50	1.06	-0.43	-0.30	-
S5	*Saya mampu merancang aktivitas belajar yang dapat mengurangi prasangka siswa akibat adanya perbedaan budaya di lingkungan belajarnya.* I am able to design learning activities that reduce prejudice in diverse classrooms.	3.71	0.98	-0.71	-0.05	-
K1	*Saya peduli terhadap keyakinan, nilai dan tradisi yang dibawa oleh setiap siswa yang berasal dari latar belakang budaya yang berbeda-beda.* I care about the beliefs, values, and traditions of students from different cultural backgrounds.	4.17	1.03	-1.29	1.12	High agreement
K2	*Saya memahami bahwa saya harus memperlakukan secara khusus setiap siswa berdasarkan latar belakang budayanya yang berbeda-beda.* I understand that students must be treated individually based on their cultural background.	3.32	1.35	-0.38	-1.05	-
K3	*Saya merasa perlu memiliki wawasan tentang berbagai gaya komunikasi siswa berdasarkan latar belakang budayanya yang berbeda-beda.* I feel the need to understand different communication styles across cultures.	4.22	1.03	-1.36	1.28	High agreement

Description: A is Awareness; S is Skill; K is Knowledge. A1 and A6 are negatively worded items and show high skewness, which may affect reliability and AVE but were retained to preserve construct coverage.

According to Table 2, most items scored moderately to high, led by the knowledge dimension (K1 = 4.17; K3 = 4.22), which implies respondents possess a solid awareness of intercultural values. However, negatively worded items A1 and A6 broke from this pattern, showing high positive skewness and kurtosis that violate normality assumptions. This uneven distribution, likely driven by item sensitivity, directly supports the need to run deeper psychometric checks via Rasch analysis and CFA. Items A1 and A6 were intentionally written as negative statements to capture the respondents' ability to reflect on possible personal bias when interacting

with students from different cultural backgrounds. The strong skewness found in these items indicates that respondents tended to reject negative attitudes, which is a common pattern in self-report instruments dealing with social and cultural sensitivity (Soto-Sanfiel et al., 2025). Such a pattern does not necessarily indicate measurement error, but rather reflects the tendency of participants to present socially acceptable responses. For this reason, the items were retained in the analysis to maintain the conceptual coverage of the awareness construct, which includes not only positive attitudes but also the ability to recognize personal prejudice (Nagy et al., 2025). In line with methodological recommendations, the presence of non-normal items was addressed by using the Maximum Likelihood Robust (MLR) estimator in the Confirmatory Factor Analysis, as this approach is considered suitable for data that deviate from normality and for samples of moderate size (Gaskin et al., 2025).

Our initial diagnostic testing revealed a strict violation of normality assumptions across all measured items (A1–A6 and S1–S3). As shown in Table 3, Shapiro-Wilk tests yielded p-values of less than 0.001 for every item, definitively rejecting the null hypothesis of univariate normality (Ghasemi and Zahediasl, 2011). We also screened for multivariate normality using Mardia's test via the MVN package in RStudio (Korkmaz et al., 2014). Establishing normality is essential here because it acts as a gating criterion for Confirmatory Factor Analysis (CFA); passing it allows researchers to use maximum likelihood estimation, which optimizes analytical accuracy (Cain et al., 2017; Yang and Liang, 2013). Since our univariate data significantly deviates from a normal curve, standard maximum likelihood estimation is not viable. Instead, the analytical strategy must shift toward robust or non-parametric estimators to properly handle the data's underlying distribution.

Table 3. Univariate Normality Test Results Using the Shapiro-Wilk Test

Item	Statistical Value	p-value	Information
A1	0.518	< 0.001	Not normally distributed
A2	0.909	< 0.001	Not normally distributed
A3	0.880	< 0.001	Not normally distributed
A4	0.889	< 0.001	Not normally distributed
A5	0.903	< 0.001	Not normally distributed
A6	0.518	< 0.001	Not normally distributed
S1	0.876	< 0.001	Not normally distributed
S2	0.897	< 0.001	Not normally distributed
S3	0.881	< 0.001	Not normally distributed
S4	0.897	< 0.001	Not normally distributed
S5	0.850	< 0.001	Not normally distributed
K1	0.769	< 0.001	Not normally distributed
K2	0.884	< 0.001	Not normally distributed
K3	0.749	< 0.001	Not normally distributed

As presented in Table 3, the Shapiro–Wilk test showed that all items had p-values below 0.001, indicating non-normal distributions across indicators and supporting the use of robust estimation procedures in the subsequent CFA analysis. Moreover, the results of the multivariate normality assessment conducted using Mardia's test revealed that the data did not conform to a normal distribution. This conclusion is supported by the Mardia skewness value of 1140.302 (p andlt; 0.001) and the Mardia kurtosis value of 10.945 (p and lt; 0.001), both of which indicate a significant deviation from the assumption of multivariate normality. The overall results of this normality test confirmed the results of the previous descriptive test, which showed deviations, although not extreme, indicating that the data did not meet the assumption of normality at both the item and aggregate item levels (Kamath et al., 2025). Such findings are common in social science research that uses Likert scales. Therefore, for further analysis of the construct validity test, the Maximum Likelihood Robust (MLR) estimation method will be used (Iacobucci et al., 2025).

Following the execution of descriptive analyses and both univariate and multivariate normality assessments, a data feasibility review was conducted using the Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy and Bartlett's Test of Sphericity. The investigation showed an overall KMO score of 0.81., categorizing it as meritorious, thereby indicating that the data were adequate for factor analysis. Additionally, the Measure of Sampling Adequacy (MSA) values for individual goods varied from 0.61 to 0.88, all exceeding the minimum threshold of 0.60, resulting in the retention of all items (Kaiser, 1974). Furthermore, Bartlett's Test of Sphericity yielded a chi-square ($\chi 2$) value of 606.10 with 91 degrees of freedom (df) and a significance level of $p < 0.001$. These findings indicate that the overall item correlations are significant and that the

correlation matrix is not an identity matrix, thus affirming the suitability of the data for subsequent analysis at the Confirmatory Factor Analysis (CFA) stage (Hair et al., 2019).

Confirmatory factor analysis (CFA) was performed using the Lavaan statistical software. The sample comprised 102 respondents. A second-order CFA was employed, featuring a construct structure with three first-order latent factors: attitude (A), skill (S), and knowledge (K), which collectively formed a second-order latent factor, namely multicultural competence. Following the consideration of descriptive tests and univariate and multivariate normality tests, the parameters were estimated using the Maximum Likelihood Robust (MLR) approach. (Yang and Liang, 2013). The model fit test results showed a chi-square ($\chi 2$) value of 140.386 with 74 degrees of freedom ($p < 0.001$), indicating a difference between the sample and model covariance matrix. The model fit test results showed a Chi-Square ($\chi 2$) value of 140.386 with 74 degrees of freedom ($p < 0.001$), indicating a difference between the sample and model covariance matrix. The robust RMSEA result of 0.089 (p-value < 0.05) indicates that the model is acceptable at the upper limit (≤ 0.10), despite the approximation breaches. Overall, the model fit test indicates that the model is reasonable, albeit it does not reach ideal criteria (Jackson et al., 2009). The following table presents the fit index results for easy inspection.

Table 4. Model Fit Test Values

Fit Index	Cutoff Criteria	Robust Value	Interpretation
χ^2 (df)	P > 0.05	97.842 (74)	Acceptable fit
CFI	≥ 0.90	0.934	Good fit
TLI	≥ 0.90	0.921	Good fit
RMSEA	≤ 0.08	0.064	Good fit

As shown in Table 4, the CFA results indicate that the proposed second-order model achieved an acceptable to good fit. Although the chi-square statistic was sensitive to sample size, the incremental fit indices met the recommended criteria (CFI = 0.934; TLI = 0.921; RMSEA = 0.064), indicating that the model fit the data well and providing empirical support for the proposed measurement structure. Meanwhile, in examining the variance and error values, it was found that most of the indicator variance errors were significant ($p < 0.05$) which indicates that although the indicators are loaded into factors (latent constructs), there is still a proportion of variance that is not explained by the factors. For example, the Skill (S) factor shows the highest reliability as reflected by the high standardized factor loading value (0.91) with a relatively small residual error (item coded S4 has a standardized variance of only 0.17), while Attitude (A) shows the lowest reliability as reflected by the low standardized factor loading (item A1 and A6 have factor loading of 0.216 and 0.290, respectively). The output of the factor loading and variance analyses is presented in the Table 5.

Table 5. Factor Loading and Variance Values

Factor	Item	Std.all	Variance Error	Variance	*p-value*
A	A1	0.216	0.953	0.203	-
	A2	0.462	0.787	0.517	0.168
	A3	0.549	0.698	0.767	0.152
	A4	0.689	0.525	0.801	0.159
	A5	0.604	0.635	0.651	0.129
	A6	0.290	0.916	0.279	0.074
S	S1	0.704	0.504	0.667	-
	S2	0.885	0.217	1,006	0,000
	S3	0.853	0.273	0.983	0,000
	S4	0.911	0.169	0.958	0,000
	S5	0.657	0.568	0.639	0,000
K	K1	0.570	0.675	0.584	-
	K2	0.504	0.746	0.678	0.019
	K3	0.719	0.482	0.736	0.012
MC	A	0.718	0.485	1,000	-
	S	0.985	0.031	0.985	0.126
	K	0.682	0.535	0.682	0.205

As presented in Table 5, the factor loading results show that the skills dimension demonstrated the strongest construct representation, with most items exhibiting high standardized loadings (0.657–0.911) and

statistically significant p-values, indicating robust indicator contributions to the latent construct. Conversely, Factor A struggled, with items like A1 and A6 standing out as weak points. Not only were their loadings statistically non-significant and low, but they also carried high error variance. Because of this underperformance, these items clearly require further refinement, an observation that strongly corroborates the Rasch findings. The Skills (S) dimension showed the strongest contribution, with standardized factor loading values between 0.657 and 0.911, reliably representing the Skills dimension. The second-order factor of Multicultural Competence also demonstrated relatively high loadings, indicating that the higher-order construct could explain the three first-order dimensions. However, the AVE values for Attitude (0.271) and Knowledge (0.344) did not reach the recommended minimum of 0.50, meaning that error variance was still larger than the variance explained by the construct. These results indicate that convergent validity has not been fully achieved, so the model should be interpreted as an initial validation of the adapted scale rather than a final measurement model, and further refinement of several indicators is still required.

Discriminant validity testing was used to determine how much dimensions empirically differ from one another, guaranteeing that each dimension uniquely measures a hidden construct. In this study, discriminant validity was determined using two methods: the Fornell–Larker criterion and the heterotrait–monotrait ratio (HTMT). The Fornell-Larcker Criterion states that a dimension has excellent discriminant validity if the square root of the average variance extracted (AVE) exceeds the correlation value between constructs. The analysis yielded square root AVE values of 0.520 for attitude (A), 0.821 for skill (S), and 0.585 for knowledge (K) dimensions. The correlation values were 0.707 between A and S, 0.490 between A and K, and 0.672 between S and K, respectively. The KM factor as a second order showed factor loading values of 0.718 to 0.985. These values confirm most dimensions-maintained discriminant validity. However, the square root AVE value of dimension A is lower than its correlation with S ($0.520 < 0.707$), indicating a potential overlap between the Attitude and Skill dimensions.

Construct reliability testing was employed to assess the internal consistency of the latent constructs (dimensions) based on the contribution of their indicators (items). In this study, construct reliability testing was conducted by examining the Composite Reliability (CR) value, with a reference value above 0.70 considered adequate (Hair et al., 2019). The findings indicate that the Skill (S) dimension exhibits the highest CR value (0.905), indicating that elements labeled S1 through S5 consistently represent the Skill dimension. The CR values for the attitude (A) and knowledge (K) dimensions were 0.648 and 0.615, respectively, which were somewhat lower than the optimal threshold. Overall, these results show that the dimensions are generally reliable and valid. However, further investigations should focus on dimensions A and K in future studies.

Meanwhile, based on the HTMT method, additional evidence was obtained by comparing the heterotrait correlation ratio (between different dimensions) with the monotrait correlation (between items that reflect the same dimension). The results showed that the pairs of *attitude* (A) and *skill* (S) dimensions had a value of 0.573, *attitude* (A) and *knowledge* (K) had a value of 0.602, and *skill* (S) and *knowledge* (K) dimensions had a value of 0.700. These values are below the threshold of 0.85, confirming that there are no serious problems related to discriminant validity between the dimensions. Thus, although dimension A contains a slight weakness in the Fornell-Larcker criterion, the combination of evidence from both methods confirms that the dimensions in the model successfully meet discriminant validity adequately. This means that the differences between the dimensions are quite clear, and there is not too much overlap.

This research was conducted as a pilot study. The psychometric analysis serves as initial validation of the multicultural competency measurement tool for prospective teachers (Villar-Guevara et al., 2024). This validation is crucial as the tool has never been validated on a large-scale sample. By conducting this validation, weaknesses in item structure, model suitability, and construct reliability can be identified (Güneşer et al., 2025). These findings can guide improvements before large-scale implementation. The data analysis includes descriptive statistics, normality tests, KMO, Bartlett's Test, model fit tests, and validity tests (Güneşer et al., 2025). The model fit test falls within the fit category, using Maximum Likelihood Robust (MLR) estimation due to non-normal distribution. For convergent validity, Skill (S) and Knowledge (K) dimensions show strong factor loadings, while Attitude (A) items remain weak (Youssef et al., 2023). The Composite Reliability test shows higher reliability in the Skill dimension compared to Attitude and Knowledge (Sovey et al., 2022). The discriminant validity test confirms each dimension measures distinct constructs.

When compared with the results of the Rasch analysis in the previous validation process, a striking difference was found. In the Rasch analysis, to meet the unidimensionality assumption, 7 of the 14 items were eliminated because they did not meet the Rasch analysis criteria (item fit) (Tesio et al., 2023). Meanwhile, the CFA test found that only 2 items in the Attitude (A) dimension had weak factor loading values, namely, values

below 0.30 (Mcneish and Wolf, 2023). Because this study was designed as a pilot study, it is not recommended to delete these items, but rather to note them as special attention for revision, both in terms of language and adaptation to the cultural context (McNeish and Wolf, 2023). Furthermore, when viewed from the second-order model fit test, it is convincingly proven that the three dimensions truly reflect the construct of multicultural competence of prospective teachers (Quast et al., 2023). This means that only the item side needs to be studied further, especially the items that reflect the Attitude (A) dimension.

In educational Unlike prior studies using either Rasch analysis or Confirmatory Factor Analysis (CFA) for validating multicultural competence instruments, this pilot study offers an integrative methodological approach. While Rasch-based studies emphasized item fit and unidimensionality, often eliminating items early, and CFA-based studies focused on latent structure without addressing measurement objectivity, this research demonstrates complementary findings (Goretzko et al., 2023). Rasch analysis identified item misfit, yet CFA supported a three-dimensional structure (Skill, Knowledge, and Attitude) with minor weaknesses in the Attitude dimension (Papini et al., 2020). The combination of Rasch and CFA outcomes in the study indicates that their combination is valid. complementary discrepancies are always complementary and they enhance the literature by showing how a second-order CFA validates. multicultural competence and Rasch analysis help to identify certain refinement needs, thereby providing an in-depth system of developing instruments (Schamberger et al., 2022).

The dissimilarity in the outcomes of Rasch examination and CFA in this pilot study are believed. to help in making the research novel. These findings underline the fact that instrument validity cannot. be defined by a single model of measurement only, rather these have to be validated by use of multiple. measurement models in order to gain a more comprehensive picture of a measurement tool (Avinç and Doğan, 2024). The objectivity of measurements is guaranteed with the help of Rasch analysis, whereas the latent structure of the is ensured. instrument is verified by use of CFA. Important methodological implications are therefore indicated, suggesting that an integrated validation strategy combining Rasch analysis and CFA should be adopted in future instrument development studies, rather than relying on a single psychometric approach. From a theoretical perspective, it is demonstrated in this research that measurement quality can only be fully understood when objectivity and construct structure are treated as complementary aspects of validity, rather than as separate concepts. For test developers and practitioners, a clear practical implication is provided: more precise, robust, and interpretable psychological instruments can be produced when both Rasch analysis and CFA are utilized during pilot and full-scale validation stages. Ultimately, a foundation is established by this pilot study for positioning the combined use of these two frameworks as a standard practice in psychological measurement development.

Two primary limitations should be taken into account when interpreting these findings. First, broader generalizability is inherently limited due to the small sample size resulting from the pilot nature of the study. Second, because deviations from normal distribution were identified in the data, the use of a Maximum Likelihood Robust estimator was required, which may have slightly affected the accuracy of the model fit. Additionally, several items within the Attitude dimension were retained despite exhibiting low factor loadings, as revisions are still needed; however, this decision may have implications for construct validity.The measurement tool needs refinement for cultural adaptation and language clarity. Future research should validate the improved instrument with a larger sample using both Rasch analysis and CFA to strengthen the multicultural competency measurement tool for prospective teachers.

CONCLUSION

Based on the series of analyses outlined above, this pilot study successfully served its purpose in the initial validation stage of the multicultural competency scale for prospective teachers before proceeding to full-scale research. The findings of the latent variable analysis using CFA show that, while the model usually passes the feasibility criteria, there are still flaws in the Attitude dimension. This discovery promotes additional investigation into the organization of elements in that dimension. Furthermore, this finding provides additional insight to the validation results using Rasch analysis in terms of the number of items successfully retained: the number of items that do not meet the CFA criteria is less than the number of items that do not meet the Rasch analysis criteria. Therefore, for subsequent research, it is recommended to change the questions in the Attitude (A) dimension and revalidate them using a larger number of respondents to ensure that the validation results are stable and universally accepted.

AUTHOR CONTRIBUTIONS

Wahyu Widodo: Conceptualization, Investigation, Data Curation, Formal Analysis, Writing - Original Draft, and Writing - Review & Editing and **Sugiantoro**: Methodology, Validation, Writing - Review & Editing, and Writing - Original Draft. All authors have read and approved the final version of this manuscript.

DATA AVAILABILITY STATEMENT

The data supporting the findings of this study are available from the authors upon reasonable request, subject to ethical approval and institutional regulations.

DECLARATION OF COMPETING INTEREST

The authors declare no known financial conflicts of interest or personal relationships that could have influenced the work reported in this manuscript.

DECLARATION OF ETHICS

The authors declare that the research and writing of this manuscript adhere to ethical standards of research and publication, in accordance with scientific principles, and are free from plagiarism.

DECLARATION OF ASSISTIVE TECHNOLOGIES IN THE WRITING PROCESS

The author state that Generative Artificial Intelligence (AI) and other assistive technologies were not excessively employed in the research and manuscript preparation process. Specifically, ChatGPT was used for translation processes and Grammarly was used for grammar and style correction. All AI-generated material was reviewed and edited for accuracy, completeness, and compliance with ethical and scholarly standards. The authors accept full responsibility for the final content of the manuscript.

REFERENCES

Afifuddin, A., Amri, M., Latif, A., Rosmini, R., and Bin Tahir, S. Z. (2025). Negotiating multicultural values within centralized education systems: A case study of Indonesia. *Frontiers in Education*, **10**, 1620685. DOI: https://doi.org/10.3389/feduc.2025.1620685.

Amarullah, R.Q., Ruslandi, R, Fadilah, R.M.Y., Ruswandi, U., and Erihadiana, M. (2024). Effective multicultural education strategies to enhance tolerance in Indonesian schools. *Atthulab: Islamic Religion Teaching and Learning Journal*, **9**(1), 142-151. DOI: https://doi.org/10.15575/ath.v9i1.28123.

Avinç, E. and Doğan, F. (2024). Digital literacy scale: Validity and reliability study with the rasch model. *Education and Information Technologies*, **29**(17), 22895–22941. DOI: https://doi.org/10.1007/s10639-024-12662-7.

Beaton, D.E., Bombardier, C., Guillemin, F., and Ferraz, M.B. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine*, **25**(24), 3186–3191. DOI: https://doi.org/10.1097/00007632-200012150-00014.

Cain, M. K., Zhang, Z., and Yuan, K.-H. (2017). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior Research Methods*, **49**(5), 1716–1735. DOI: https://doi.org/10.3758/s13428-016-0814-1.

Erdem, D. (2020). Multicultural competence scale for prospective teachers: Development, validation and measurement invariance. *Eurasian Journal of Educational Research*, **20**(87), 1–28. Retrieved from: https://ejer.com.tr/multicultural-competence-scale-for-prospective-teachers-development-validation-and-measurement-invariance/.

Fatmawati, L., Dewi, K.P., and Wuryandani, W. (2023). Multicultural competence of elementary teacher education students. *International Journal of Elementary Education*, **7**(4), 721–730. DOI: https://doi.org/10.23887/ijee.v7i4.62880.

Fornell, C. and Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, **18**(1), 39–50. DOI: https://doi.org/10.1177/002224378101800104.

Gaskin, J.E., Lowry, P.B., Rosengren, W., and Fife, P.T. (2025). Essential validation criteria for rigorous covariance-based structural equation modelling. *Information Systems Journal*, **35**(6), 1630–1661. DOI: https://doi.org/10.1111/isj.12598.

Ghasemi, A. and Zahediasl, S. (2011). Normality tests for statistical analysis: A guide for non-statisticians. *International Journal of Endocrinology and Metabolism*, **10**(2), 486-489. DOI: https://doi.org/10.5812/ijem.3505.

Goretzko, D., Siemund, K., and Sterner, P. (2023). Evaluating model fit of measurement models in confirmatory factor analysis. *Educational and Psychological Measurement*, **84**(1), 123–144. DOI: https://doi.org/10.1177/00131644231163813.

Güneşer, R., Kirimlioğlu, N., and Kalaycioğlu, O. (2025). Validity and reliability of the Turkish version of the Ethical sensitivity scale. *Ethics and Behavior*, 1–14. DOI: https://doi.org/10.1080/10508422.2025.2534938.

Hair, J.F. (2019). *Multivariate data analysis* (Eighth edition). Boston: Cengage.

Hair, J.F., Black, W.C., Babin, B.J., and Anderson, R.E. (2019). *Multivariate data analysis* (Eighth edition). Boston: Cengage.

Harbison, L. and Simmons, K. (2024). Fundamentals of descriptive statistics. *American Dental Hygienists' Association*, **98**(5), 51–54. Retrieved from: https://pubmed.ncbi.nlm.nih.gov/39406491/.

Henseler, J., Ringle, C.M., and Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science*, **43**(1), 115–135. DOI: https://doi.org/10.1007/s11747-014-0403-8.

Iacobucci, D., Román, S., Moon, S., and Rouziès, D. (2025). A tutorial on what to do with skewness, kurtosis, and outliers: New insights to help scholars conduct and defend their research. *Psychology and Marketing*, **42**(5), 1398–1414. DOI: https://doi.org/10.1002/mar.22187.

Indu, P.V., Vidhukumar, K., Chacko, D., Menon, V., Grover, S., and Gupta, S. (2025). Criterion validity, construct validity, and factor analysis: An introductory overview. *Indian Journal of Psychiatry*, **67**, 916-921. DOI: https://doi.org/10.4103/indianjpsychiatry_911_25.

Jackson, D.L., Gillaspy, J.A., and Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods*, **14**(1), 6–23. DOI: https://doi.org/10.1037/a0014694.

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

Kamath, A., Poojari, S., and Varsha, K. (2025). Assessing the robustness of normality tests under varying skewness and kurtosis: a practical checklist for public health researchers. *BMC Medical Research Methodology*, **25**(1), 206. DOI: https://doi.org/10.1186/s12874-025-02641-y.

Khalim, A.D.N. and Parut, W. (2025). Paradigma and programs multicultural education in inclusive madrasah. *Journal of Contemporary Islamic Education*, **5**(1), 28–44. DOI: https://doi.org/10.25217/jcie.v5i1.5115.

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

Korkmaz, S., Goksuluk, D., and Zararsiz, G. (2014). MVN: An R package for assessing multivariate normality. *The R Journal*, **6**(2), 151-162. DOI: https://doi.org/10.32614/RJ-2014-031.

Mariyono, D. (2024). Indonesian mosaic: The essential need for multicultural education. *Quality Education for All*, **1**(1), 301–325. DOI: https://doi.org/10.1108/QEA-05-2024-0042.

Mcneish, D. and Wolf, M.G. (2023). Dynamic fit index cutoffs for confirmatory factor analysis models. *Psychological Methods*, **28**(1), 61–88. DOI: https://doi.org/10.1037/met0000425.

Nagy, T., Hergert, J., and Elsherif, M.M. (2025). Bestiary of questionable research practices in psychology. *Advances in Methods and Practices in Psychological Science*, **8**(3), 25152459251348431. DOI: https://doi.org/10.1177/25152459251348431.

Ningtyas, D.W. (2024). Peran guru dalam pendidikan multikultural untuk membentuk karakter peserta didik di era society 5.0. *Jurnal MIPA dan Pembelajarannya*, **4**(3), 3. Retrieved from: https://journal3.um.ac.id/index.php/mipa/article/view/5306.

Papini, N., Kang, M., Ryu, S., Griese, E., Wingert, T., and Herrmann, S. (2020). Rasch calibration of the 25-item Connor-Davidson Resilience Scale. *Journal of Health Psychology*, **26**(11), 1976–1987. DOI: https://doi.org/10.1177/1359105320904769.

Pratama, R., Sumantri, S.H., and Widodo, P. (2023). The role of teachers in implementing multicultural education at Taruna Nusantara High School to enhance social resilience. *International Journal of Humanities Education and Social Sciences (IJHESS)*, **3**(1), 167-176. DOI: https://doi.org/10.55227/ijhess.v3i1.580.

Pratiwi, H., Dwiningrum, S.I.A., Riwanda, A., and Minasyan, S. (2024). Insights into multicultural competence of early childhood teacher candidates in Indonesian Islamic Higher Education. *EDUKASI: Jurnal Penelitian Pendidikan Agama dan Keagamaan*, **22**(1), 79–96. DOI: https://doi.org/10.32729/edukasi.v22i1.1813.

Quast, J., Rubach, C., and Porsch, R. (2023). Professional digital competence beliefs of student teachers, pre-service teachers and teachers: Validating an instrument based on the DigCompEdu framework. *European Journal of Teacher Education*, **48**(4), 698–721. DOI: https://doi.org/10.1080/02619768.2023.2251663.

Schamberger, T., Schuberth, F., and Henseler, J. (2022). Confirmatory composite analysis in human development research. *International Journal of Behavioral Development*, **47**(1), 89–100. DOI: https://doi.org/10.1177/01650254221117506.

Soto-Sanfiel, M.T., Angulo-Brunet, A. and Lutz, C. (2025). The scale of artificial intelligence literacy for all (SAIL4ALL): Assessing knowledge of artificial intelligence in all adult populations. *Humanities and Social Sciences Communications*, **12**, 1618. DOI: https://doi.org/10.1057/s41599-025-05978-3.

Sovey, S., Osman, K., and Mohd-Matore, M.E.E. (2022). Exploratory and confirmatory factor analysis for disposition levels of computational thinking instrument among secondary school students. *European Journal of Educational Research*, **11**(2), 639–652. DOI: https://doi.org/10.12973/eu-jer.11.2.639.

Spanierman, L.B., Oh, E., Heppner, P.P., Neville, H.A., Mobley, M., Wright, C.V., Dillon, F.R., and Navarro, R. (2011). The multicultural teaching competency scale: development and initial validation. *Urban Education*, **46**(3), 440–464. DOI: https://doi.org/10.1177/0042085910377442.

Tesio, L., Caronni, A., Simone, A., Kumbhare, D., and Scarano, S. (2023). Interpreting results from Rasch analysis 2. Advanced model applications and the data-model fit assessment. *Disability and Rehabilitation*, **46**(3), 604–617. DOI: https://doi.org/10.1080/09638288.2023.2169772.

Umar, J. and Nisa, Y.F. (2020). Uji validitas konstruk dengan CFA dan pelaporannya. *Jurnal Pengukuran Psikologi dan Pendidikan Indonesia (JP3I)*, **9**(2), 1–11. DOI: https://doi.org/10.15408/jp3i.v9i2.16964.

Villar-Guevara, M., Livia-Segovia, J.H., García-Salirrosas, E.E., and Fernández-Mallma, I. (2024). Student Evaluation of Teachers' Effectiveness (SETE) scale: Translation, cross-cultural adaptation and psychometric properties in a Latin American sample. *Frontiers in Education*, **9**, 1401718. DOI: https://doi.org/10.3389/feduc.2024.1401718.

Widodo, W. and Chotimah, C. (2023). Adaptasi dan analisis psikometri skala kompetensi multikultural calon guru menggunakan pemodelan Rasch. *Jurnal Pendidikan dan Kebudayaan*, **8**(2), 153–172. DOI: https://doi.org/10.24832/jpnk.v8i2.4228.

Yang, Y. and Liang, X. (2013). Confirmatory factor analysis under violations of distributional and structural assumptions. *International Journal of Quantitative Research in Education*, **1**(1), 61–84. DOI: https://doi.org/10.1504/IJQRE.2013.055642.

Youssef, N., Saleeb, M., Gebreal, A., and Ghazy, R.M. (2023). The internal reliability and construct validity of the Evidence-Based Practice Questionnaire (EBPQ): Evidence from healthcare professionals in the eastern mediterranean region. *Healthcare*, **11**(15), 2168. DOI: https://doi.org/10.3390/healthcare11152168.