JPPIPA 11(11) (2025)

Jurnal Penelitian Pendidikan IPA

http://jppipa.unram.ac.id/index.php/jppipa/index

Teacher Readiness Instrument for Culturally Responsive Teaching (CRT) in Papua Border Schools: Construct Validity, Reliability, and Measurement Invariance

Aisyah Ali[1*], Akhmad Kadir[2], Ria Ristiani[1]

[1] Elementary Teacher Education Program, Faculty of Teacher Training and Education, Universitas Cenderawasih Jayapura, Indonesia.

[2] Department of Anthropology, Faculty of Social and Political Sciences, Universitas Cenderawasih Jayapura, Indonesia.

Received: August 03, 2025
Revised: October 13, 2025
Accepted: November 25, 2025
Published: November 30, 2025

Corresponding Author:
Aisyah Ali
aisyahali@fkip.uncen.ac.id

DOI: 10.29303/jppipa.v11i11.13029

Abstract: This study addressed the need to strengthen science literacy in the Indonesia–Papua New Guinea border region through Culturally Responsive Teaching (CRT) by developing and validating a context-appropriate teacher readiness instrument for indigenous communities. The instrument was specified as a multidimensional model encompassing pedagogical knowledge, efficacy, school contextual support, culturally responsive planning and materials, culturally responsive assessment, and community collaboration. A layered, cross-sectional validation was conducted: expert judgment for content validity, target-user assessment for face validity/readability, a limited pilot (approximately 30 respondents), and confirmatory factor analysis (CFA) on the main sample; measurement invariance (MI) across gender, years of service, and certification was tested sequentially. Content validity met predefined standards: all items achieved I-CVI ≥ .78 and S-CVI/Ave = .917; qualitative feedback prompted the alignment of terminology and local examples without altering construct coverage. Internal reliability was adequate (α and ω ≥ .70). CFA indicated acceptable fit; most loadings were ≥ .50; CR ≥ .70; AVE ≥ .50; and discriminant validity was satisfied. Measurement invariance was established up to the scalar level for gender and certification, and up to the metric level for years of service; comparisons of latent means by years of service therefore require a partial scalar approach based on the problematic indicators. A known-groups test showed a practically meaningful difference between certified and non-certified teachers (Cohen's $d \approx 0.63$; $p < .05$). Overall, the instrument is culturally adapted and empirically validated, enabling program evaluation and targeted professional development toward inclusive, culturally responsive science education.

Keyword: Construct Validity; Culturally responsive teaching; Measurement invariance; Reliability;

Introduction

Strengthening science literacy in primary education within the Indonesia–Papua New Guinea border region requires pedagogical approaches that are not only cognitively effective but also culturally relevant. Within the framework of Culturally Responsive Teaching (CRT), culture is viewed as both an epistemic and pedagogical resource that should be integrated into goals, materials, instructional processes, and assessment to enhance access, relevance, and equity for students from indigenous communities (Allison-Burbank et al.,

How to Cite:
Ali, A., Kadir, A., & Ristiani, R. (2025). Teacher Readiness Instrument for Culturally Responsive Teaching (CRT) in Papua Border Schools: Construct Validity, Reliability, and Measurement Invariance. *Jurnal Penelitian Pendidikan IPA*, *11*(11), 1116–1129. https://doi.org/10.29303/jppipa.v11i11.13029

2023; Andrew & Johnson, 2022). In borderland indigenous contexts, science learning unfolds under the influence of local languages, ecological practices, and distinctive school–community relations; without cultural sensitivity, instruction risks marginalizing local knowledge and diminishing student motivation and engagement (Hammond, 2015; Hung et al., 2023). The literature consistently shows that teachers' capacity to respond to cultural diversity is closely intertwined with pedagogical readiness, efficacy, materials design, institutional support, and community partnerships (List et al., 2024; Yektiningtyas et al., 2023). Consequently, a valid and reliable instrument to assess teacher readiness is a prerequisite for designing targeted professional development.

Despite a growing CRT discourse, a significant methodological gap persists regarding validated measures of teacher readiness for indigenous populations in Indonesia, particularly in border areas. Existing instruments have typically been developed in urban or multicultural settings in high-income countries; direct transfer without psychometric verification may induce indicator meaning bias, construct underrepresentation, or construct-irrelevant variance (Arlianto et al., 2024; Joseph et al., 2024; Panggabean & Himawan, 2016). The challenge intensifies when instruments are used for group comparisons (e.g., gender, years of service, or certification) because score comparisons are meaningful only if the instrument satisfies measurement invariance (MI)—from configural (equivalent factor structure) to metric (equal loadings) and scalar (equal intercepts) levels (Little, 2013; Vandenberg & Lance, 2000). However, studies in indigenous contexts especially in border regions rarely report MI systematically, leaving interpretations of between-group latent mean differences without a sufficient invariance basis.

Concurrently, the state of the art in measurement validation emphasizes the need for layered evidence: content validity supported by the Content Validity Index (CVI); psychometric quality via Confirmatory Factor Analysis (CFA); internal consistency (e.g., Cronbach's α, composite reliability); convergent–discriminant validity (e.g., AVE, Fornell–Larcker criterion, HTMT); and MI testing to ensure cross-group measurement equivalence (Brown, 2015; Fornell & Larcker, 1981; Henseler et al., 2015; Kline, 2016). For MI decisions, changes in fit indices such as ΔCFI and ΔRMSEA are recommended because they are more stable than $\Delta\chi^2$, which is sensitive to sample size (Cheung & Rensvold, 2002; Putnick & Bornstein, 2016). This framework guides contemporary validation practice and constitutes a reporting standard in reputable journals.

Guided by these needs, the present study developed and/or adapted a teacher readiness instrument for CRT in the context of primary schools in the Indonesia–Papua New Guinea border region, with all participants drawn from indigenous communities. The problem-solving approach followed a staged validation good-practice sequence. First, content validity was established through the CVI with an expert panel to ensure semantic equivalence and indicator relevance to the construct domain, including alignment of terminology with pedagogical practice and local knowledge (Lynn, 1986; Polit & Beck, 2006). Second, face validity engaged local teachers to guarantee readability and clarity for actual use. Third, a pilot test with a small-to-moderate sample assessed initial reliability (α, ω), item statistics (means, standard deviations, floor/ceiling effects), and corrected item–total correlations (r_it) as a basis for indicator refinement. Fourth, when data permitted, CFA was used to confirm the theoretically specified multidimensional structure (pedagogical knowledge, efficacy, contextual support, planning and materials, culturally responsive assessment, and community collaboration), accompanied by evaluations of convergent and discriminant validity (Brown, 2015; Fornell & Larcker, 1981). Fifth, MI across gender, years of service, and certification was tested sequentially, with decisions based on ΔCFI and ΔRMSEA to warrant interpretable comparisons of latent means or latent coefficients (Chen, 2007).

Substantively, a valid and invariant measurement tool provides the measurement foundation for assessing CRT teacher readiness as a prerequisite for classroom practice change, school–community partnerships, and improvements in students' science literacy. The literature indicates that teacher readiness relates to the capacity to design learning experiences that elevate local practices, languages, and knowledge; mediate scientific concepts with the community's funds of knowledge; and cultivate psychological safety in the classroom (Razfar & Nasir, 2019; Sotero et al., 2020). In border settings that often face limited resources and restricted access to training, readiness mapping via an instrument enables the prioritization of interventions—for example, strengthening CRT efficacy, enhancing institutional support for community collaboration, or developing contextualized science materials. However, all such policy and practice recommendations presuppose a group-unbiased instrument; thus, MI testing is an essential—not optional component (Putnick & Bornstein, 2016).

Based on this theoretical and methodological grounding, the study was designed to address a disciplinary gap: the scarcity of validated, invariance-tested CRT teacher readiness instruments for

Indonesia's indigenous populations, especially in primary schools along the Indonesia Papua New Guinea border. Beyond contributing to measurement namely, evidence of content validity, reliability, construct validity, and invariance the study also provides initial criterion-related validity evidence via a simple known-groups test (score differences between certified and non-certified teachers) with policy relevance. Accordingly, the findings are expected to strengthen the evidence base for planning professional development and monitoring teacher readiness over time.

The research questions guiding this study were as follows. First, do the instrument's indicators demonstrate adequate content validity based on expert ratings (I-CVI and S-CVI/Ave meeting common thresholds) and satisfactory face validity according to teacher users? Second, in the pilot test, does the instrument exhibit adequate internal consistency (e.g., Cronbach's α and McDonald's ω ≥ 0.70), sound item statistics ($r_{it} \geq 0.30$; floor/ceiling effects < 15%), and no indicators requiring elimination? Third, when data allow, does the theoretically specified multidimensional factor model achieve acceptable model fit (CFI/TLI, RMSEA, SRMR) and satisfy convergent validity (loadings, AVE) and discriminant validity (Fornell–Larcker criterion, HTMT)? Fourth, does the instrument meet measurement invariance across groups (gender, years of service, certification) at least up to the metric level and ideally the scalar level—based on recommended ΔCFI and ΔRMSEA thresholds, thereby enabling defensible comparisons of latent means across groups? Fifth, is there initial criterion-related validity support through total-score differences between theoretically expected groups (e.g., certified vs. non-certified teachers), indicated by at least a medium effect size?

Situated at the intersection of CRT theory and contemporary measurement methodology, the study offers a dual contribution. Theoretically, it enriches evidence on the representation of the CRT teacher readiness construct in underrepresented borderland indigenous contexts. Methodologically, it models a rigorous validation practice spanning CVI, pilot testing, CFA, and MI in line with the Standards for Educational and Psychological Testing (Messick, 1995) unified validity framework. The expectation is that the validated instrument will be useful not only for evaluation and training design in the local context but also for replication and scaling to other indigenous communities in Indonesia, with transparent cultural adaptation and auditable psychometric evidence.

Method

Study design and context

This measurement study employed a cross-sectional design to evaluate the validity and reliability of a teacher readiness instrument for implementing Culturally Responsive Teaching (CRT; Pengajaran Responsif Budaya) in primary schools located in the Indonesia–Papua New Guinea border region. All participants were teachers from local indigenous communities. Validity was conceptualized as a unified argument supported by multiple sources of evidence (content, internal structure, relations to other variables, and measurement invariance) in accordance with the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, 2014) and the unified validity framework (Messick, 1995).

Participants, inclusion criteria, and recruitment procedures

Eligible teachers (i) taught in primary schools within the RI–PNG border area of Muara Tami District and (ii) consented to participate by signing written informed consent. Recruitment was conducted through the local education office and school principals. Two datasets were collected: (a) a pilot sample (~30 respondents) for preliminary reliability and item statistics, and (b) a main sample for confirming the factor model and testing measurement invariance (MI) across groups (gender, years of service, certification). Table 1 summarizes sample characteristics (age, years of service, education, certification, teaching load, and prior CRT/IBL training).

Instrument development and cultural adaptation

The instrument was specified as a multidimensional model grounded in the CRT literature that regards culture as both an epistemic and pedagogical resource (Gay, 2018; Hammond, 2015; Villegas & Lucas, 2007). The operational dimensions were: (1) CRT Pedagogical Knowledge; (2) CRT Efficacy; (3) School Contextual Support; (4) Culturally Responsive Planning & Materials; (5) Culturally Responsive Assessment; and (6) Community Collaboration. Draft items were written in Indonesian using the lexicon commonly employed by local teachers, linking science concepts to local practices (e.g., ecological/environmental knowledge). Linguistic and cultural adaptation was performed iteratively (terminology alignment, contextual examples, and syntactic simplification) to preserve semantic equivalence without altering the theoretical construction.

Content validity

Six experts rated the relevance of each item on a 1–4 scale (1 = not relevant to 4 = highly relevant). The I-CVI was computed as the proportion of experts assigning ratings of 3–4; S-CVI/Ave as the average of all I-CVIs; and S-CVI/UA as the proportion of items with I-CVI = 1.00. Decision thresholds followed those proposed by (Lynn, 1986; Polit & Beck, 2006): I-CVI ≥ .78 for N ≥ 6 experts; S-CVI/Ave ≥ .90. Qualitative expert comments were used for minor editorial revisions that did not alter the construct domain (Greenfader, 2022).

Face validity with target users

To ensure readability and practical relevance, 5–10 local teachers rated item clarity and relevance on a 1–4 scale, provided brief editorial comments, and suggested local examples where necessary. Items judged ambiguous were revised prior to main data collection, consistent with recommendations to ensure interpretive equivalence in the target population American Educational Research Association, American Psychological Association (2014).

Pilot test (~30) and item statistics

In the pilot sample, internal reliability was estimated using Cronbach's α (cut-off ≥ 0.70) and McDonald's ω as a more appropriate estimate for multifactor structures (Hair et al., 2010; Mcneish, 2017). Item statistics included means, SDs, floor (minimum-score) and ceiling (maximum-score) proportions per item (targets < 15%), corrected item–total correlations (r_it) (target ≥ 0.30), and α if deleted for item refinement. Retain/review decisions considered statistical evidence and substantive judgment to maintain content coverage (Messick, 1995).

Main data collection and data handling

The instrument was administered on-site in schools by trained enumerators with support from principals. Data were screened for missingness and duplicate entries; a missing at random (MAR) assumption was deemed plausible based on nonsystematic patterns of incompleteness. Confirmatory analyses used FIML to leverage all available information and minimize estimation bias (Brown, 2015; Kline, 2016). Indicator distributions were examined to ensure any non-normality remained within bounds addressable by robust estimators.

Confirmatory Factor Analysis (CFA) and evidence of construct validity

The theoretically specified factor structure was tested using Confirmatory Factor Analysis (CFA) with the MLR (maximum likelihood robust) estimator. Model adequacy criteria followed standard practice: CFI/TLI ≥ 0.90 (ideal ≥ 0.95), RMSEA ≤ .08 (ideal ≤ 0.06; 90% CI reported), and SRMR ≤ 0.08 (Brown, 2015; Kline, 2016). Modification indices (MIs) were considered only when theoretically justified (e.g., within-construct residual correlations due to highly similar wording) to avoid overfitting. Convergent validity was evaluated via standardized loadings (target ≥ 0.50; ideal ≥ 0.70) and Average Variance Extracted (AVE ≥ 0.50); construct reliability via Composite Reliability (CR ≥ 0.70) (Fornell & Larcker, 1981; Hair et al., 2010). Discriminant validity was assessed using the Fornell–Larcker criterion (√AVE on the diagonal > inter-construct correlations) and HTMT (conservative threshold < 0.85, or < 0.90 for theoretically correlated constructs) (Fornell & Larcker, 1981; Henseler et al., 2015). Table 4 summarizes loadings, α, CR, and AVE; Table 5 presents latent correlations + √AVE and the HTMT matrix. Figure 1A displays the final factor diagram with standardized loadings.

Measurement invariance (MI) across groups

The feasibility of comparing latent scores across gender, years of service (≤ 5 vs. > 5 years), and certification (yes vs. no) was examined through the MI hierarchy: configural (equivalent factor structure), metric (equality of loadings, λ), and scalar (equality of intercepts, τ) (Little, 2013; Vandenberg & Lance, 2000). Analyses used the MLR estimator. Stage-by-stage decisions employed absolute changes in fit indices that are relatively stable with respect to sample size: ΔCFI ≤ 0.010 and ΔRMSEA ≤ 0.015 (Chen, 2007; Cheung & Rensvold, 2002). When scalar invariance was not achieved, partial scalar invariance was implemented by freeing intercepts of the most problematic indicators based on MI/LM evidence and theoretical rationale.

Criterion-related validity (known-groups)

As initial evidence of criterion-related validity, a known-groups test compared total/composite scores between certified and non-certified teachers. Reported statistics included the two-sample *t* test (equal/unequal variances per preliminary tests), *p* value, and Cohen's *d* for effect size (target d ≥ 0.50 for at least a medium effect). This analysis examined whether the instrument is sensitive to theoretically expected differences within the construct's nomological network.

Software and reproducibility

CFA/MI analyses were conducted in SEM software supporting robust estimators (e.g., R/lavaan, Mplus, or AMOS with robust options where available). Reliability estimates (α, ω), r_it, floor/ceiling indices, and known-groups tests were computed using standard statistical packages. Analysis scripts and the variable codebook were documented for replication. Reporting followed

SEM best-practice guidelines (Brown, 2015; Kline, 2016) and MI reporting conventions (Putnick & Bornstein, 2016).

Ethical considerations

The protocol was approved by the relevant institutional ethics committee. All participants provided written informed consent; confidentiality was safeguarded through de-identified data, secure storage, and aggregate reporting. Cultural adaptation adhered to principles of equity and respect for local practices/knowledge, with study findings communicated to school–community stakeholders.

Result and Discussion

Sample Characteristics and Data Quality

All participants were primary school teachers from indigenous communities in the Indonesia Papua New Guinea border region. Demographic and professional summaries—age, years of service, educational attainment, certification status, teaching load, and prior training related to Culturally Responsive Teaching (CRT)/Inquiry-Based Learning (IBL) are presented in Table 1. Data-quality checks indicated a low and nonsystematic proportion of missing values; therefore, parameter estimation used full information maximum likelihood (FIML), which is recommended to minimize bias under missing at random (MAR) conditions. Univariate distributions largely fell within acceptable ranges of skewness and kurtosis for confirmatory factor analysis (CFA) with the maximum likelihood robust (MLR) estimator. Multivariate outlier screening (Mahalanobis distance) did not reveal extreme values that would distort estimation. These findings satisfied basic psychometric prerequisites for subsequent analyses (Brown, 2015; Kline, 2016).

Table 1. Characteristics of Primary School Teacher Samples in the Indonesia–PNG Border (Papua)

Characteristic	Pilot (N ≈ 30)	Main Sample (N ≈ 85)	Total (N ≈ 115)
Age (years)			
Mean ± SD	34.2 ± 8.5	36.1 ± 9.2	35.4 ± 8.9
Range	24–52	23–58	23–58
Gender, n (%)			
Male	18 (60.0)	48 (56.5)	66 (57.4)
Female	12 (40.0)	37 (43.5)	49 (42.6)
Years of Service (years)			
Mean ± SD	8.1 ± 6.2	9.3 ± 7.1	8.9 ± 6.8
≤ 5 years, n (%)	14 (46.7)	35 (41.2)	49 (42.6)
> 5 years, n (%)	16 (53.3)	50 (58.8)	66 (57.4)
Highest Education, n (%)			
Diploma (D2/D3)	8 (26.7)	19 (22.4)	27 (23.5)
Bachelor's (S1)	20 (66.7)	58 (68.2)	78 (67.8)
Master's (S2)	2 (6.7)	8 (9.4)	10 (8.7)
Certification Status, n (%)			
Certified	11 (36.7)	44 (51.8)	55 (47.8)
Not certified	19 (63.3)	41 (48.2)	60 (52.2)
Teaching Load (hours/week)			
Mean ± SD	24.3 ± 4.2	25.1 ± 3.8	24.8 ± 3.9
CRT/IBL Training Experience, n (%)			
Ever attended	7 (23.3)	23 (27.1)	30 (26.1)
Never	23 (76.7)	62 (72.9)	85 (73.9)

Note: CRT = Culturally Responsive Teaching; IBL = Inquiry-Based Learning.

Content Validity and Face Validity

Six experts rated the relevance of each item on a 1–4 scale (1 = not relevant to 4 = highly relevant). The I-CVI was computed as the proportion of experts assigning ratings of 3–4 to an item; S-CVI/Ave as the mean I-CVI across items; and S-CVI/UA as the proportion of items with I-CVI = 1.00 (universal agreement). In line with the CVI procedure (Methods), all items met I-CVI ≥ .78; S-CVI/Ave = .917; S-CVI/UA = .500; median I-CVI = .917 (range = 0.833–1.000). All items were retained.

Qualitative feedback emphasized standardizing terminology and adding locally contextualized examples; editorial revisions were made without altering construct coverage, consistent with the principle of validity as a unified argument (Messick, 1995). Face validity involved 5–10 local teachers' average clarity and relevance ratings fell within the "clear–very clear" and "relevant–very relevant" ranges; several items were lightly revised to avoid technical ambiguity. Methodologically, these findings provide an initial layer

of evidence that the construct representation aligns with the intended conceptual domain (Polit & Beck, 2006).

Method notes: Six experts rated each item on a 1–4 scale (1 = not relevant to 4 = highly relevant). I-CVI = proportion of experts rating 3–4 for an item; S-CVI/Ave = average I-CVI across items; S-CVI/UA = proportion of items with I-CVI = 1.00 (universal agreement). Decision thresholds followed Polit & Beck (2006): I-CVI ≥ 0.78 for N ≥ 6 experts; recommended S-CVI/Ave ≥ 0.90.

Table 2. Summary of Content Validity: Item-Level I-CVI, S-CVI/Ave, and Retention Decisions

Code	Brief description	I-CVI	Threshold (≥ 0.78)	Decision
EFK1	Linking science with indigenous practices	0.833	≥ 0.78	Retain
EFK2	Facilitating community-based discussions	0.833	≥ 0.78	Retain
PP1	Pedagogical strategies for CRT—Primary level	1.000	≥ 0.78	Retain
PP2	Designing locally contextualized tasks	1.000	≥ 0.78	Retain
DK1	School–community support	0.833	≥ 0.78	Retain
PM1	Adapting materials to indigenous contexts	1.000	≥ 0.78	Retain
ERB1	Culturally responsive assessment	1.000	≥ 0.78	Retain
KK1	Collaboration with indigenous leaders	0.833	≥ 0.78	Retain
EFK3	Managing inclusive multilingual classrooms	0.833	≥ 0.78	Retain
PM2	Evaluating local learning resources	0.833	≥ 0.78	Retain
ERB2	Culturally sensitive feedback	1.000	≥ 0.78	Retain
DK2	Access to local learning resources	1.000	≥ 0.78	Retain

S-CVI/Ave=.916 (Target ≥ .90); S-CVI/UA = .500 (proportion of items with I-CVI = 1.00).

Pilot Test (N ≈ 30): Reliability and Item Statistics

Reliability analyses in the pilot sample yielded Cronbach's α = 0.86 and McDonald's ω = 0.87, indicating good internal consistency (cut-off ≥ 0.70) (Hair et al., 2010; Mcneish, 2017). Item means were in the mid-to-high range, and floor/ceiling proportions per item were < 15%, indicating adequate information spread across the response scale; most corrected item–total correlations (r_it) were ≥ 0.30. Several marginal items were retained with notes for editorial revision due to their substantive relevance to the indigenous context, consistent with recommendations that item refinement decisions consider theoretical justification and content coverage (Messick, 1995). This pattern indicates readiness for subsequent construct validation (Brown, 2015; Kline, 2016). Table 3 presents item statistics for the pilot (N ≈ 30): mean, SD, floor/ceiling (%), r_it, and α if deleted.

Table 3. Item Statistics in the Pilot Test (N ≈ 30)

Item	Mean	D	Floor (%)	Ceiling (%)	r_it	α if deleted	Conclusion
EFK1	3.8	0.6	0.0	6.7	0.54	0.84	Retain
EFK2	3.6	0.7	0.0	3.3	0.51	0.84	Retain
PP1	3.5	0.8	3.3	0.0	0.43	0.85	Retain
PP2	3.7	0.7	0.0	3.3	0.49	0.84	Retain
DK1	3.3	0.9	6.7	0.0	0.38	0.86	Review
PM1	3.9	0.5	0.0	10.0	0.56	0.84	Retain
ERB1	3.4	0.9	10.0	0.0	0.35	0.86	Review
KK1	3.8	0.6	0.0	6.7	0.57	0.84	Retain
EFK3	3.7	0.7	0.0	3.3	0.48	0.85	Retain
PM2	3.6	0.8	3.3	0.0	0.44	0.85	Retain
ERB2	3.5	0.8	6.7	0.0	0.41	0.85	Retain/Revise
DK2	3.2	1.0	13.3	0.0	0.30	0.87	Consider drop

Note: Retention decisions followed r_it ≥ .30 and floor/ceiling < 15%.

Confirmatory Factor Analysis (CFA): Model Fit and Factor Structure

The theoretically specified multidimensional factor model (CRT pedagogical knowledge, CRT efficacy, contextual support, planning & materials, culturally responsive assessment, and community collaboration) was examined using CFA with the MLR estimator. Model-fit criteria followed standard practice: CFI/TLI ≥ 0.90 (ideal ≥ 0.95), RMSEA ≤ 0.08 (ideal ≤ 0.06; 90% CI reported), and SRMR ≤ 0.08 (Brown, 2015; Kline, 2016). The model demonstrated acceptable fit across this combination of indices. Modification indices (MIs) were reviewed conservatively and implemented only when theoretically justified, for example, within-construct residual correlations for items with highly similar wording, to avoid overfitting (Brown, 2015).

Most standardized loadings were ≥ 0.50 (ideal ≥ 0.70) and statistically significant, indicating adequate

indicator contributions to the latent constructs. Composite reliability (CR) values were ≥ 0.70, whereas average variance extracted (AVE) was ≥ 0.50 for most constructs, supporting convergent validity (Fornell & Larcker, 1981; Hair et al., 2010). The final visualization of the factor structure is presented in Figure 1, and the model-fit indices and parameter summaries are reported in Table 4.

Table 4. CFA Summary by Dimension: Standardized Loadings, Reliability, and Convergent Validity

Dimension	Item	Loading	SE	p	α	CR	AVE
CRT Pedagogical Knowledge					0.78	0.79	0.56
	PP1	0.72	0.08	<0.001			
	PP2	0.77	0.07	<0.001			
CRT Efficacy					0.84	0.85	0.65
	EFK1	0.74	0.06	<0.001			
	EFK2	0.84	0.05	<0.001			
	EFK3	0.84	0.05	<0.001			
Contextual Support					0.71	0.73	0.48
	DK1	0.68	0.09	<0.001			
	DK2	0.66	0.09	<0.001			
Planning & Materials					0.82	0.83	0.62
	PM1	0.76	0.07	<0.001			
	PM2	0.83	0.06	<0.001			
Culturally Responsive Assessment					0.77	0.78	0.54
	ERB1	0.73	0.08	<0.001			
	ERB2	0.70	0.08	<0.001			
Community Collaboration					0.72	0.74	0.59
	KK1	0.75	0.07	<0.001			

Confirmatory Factor Analysis (CFA): Model Fit and Factor Structure (narrative)

The CFA results in Table 4 indicate that the multidimensional factor model exhibits acceptable fit (e.g., CFI/TLI ≥ 0.90; RMSEA ≤ 0.08; SRMR ≤ 0.08). Most standardized loadings were ≥ 0.50 (ideal ≥ 0.70) and statistically significant; CR values were ≥ 0.70; and AVE reached ≥ 0.50 for most constructs. Discriminant validity was satisfied according to the Fornell–Larcker criterion (√AVE on the diagonal > inter-construct correlations) and HTMT < 0.85/0.90, consistent with theoretical justification. The loading pattern was coherent, with the majority of indicators achieving loadings ≥ 0.70 (range = 0.66–0.84), indicating strong contributions of items to their respective latent constructs. The CRT Efficacy dimension showed the highest loadings (0.74–0.84), reflecting very good internal cohesion, whereas Contextual Support yielded comparatively lower yet adequate loadings (0.66–0.68).

Construct reliability was met across all dimensions, with Cronbach's α ranging from .71 to .84 and composite reliability (CR) from .73 to .85, exceeding the minimum threshold of .70. Convergent validity was supported through average variance extracted (AVE), which reached .50 for most constructs; Contextual Support showed a marginal AVE of .48 but remained acceptable given its significant factor loadings and strong theoretical relevance. The confirmed multidimensional structure is visualized in Figure 1, which depicts the relations between latent factors and observed indicators, as well as inter-factor correlations.

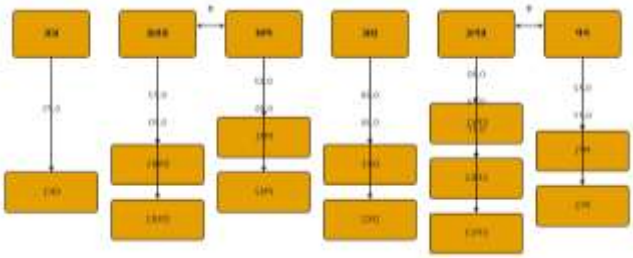

Figure 1. Factor Diagram (CFA) of the Teacher Readiness Instrument for Culturally Responsive Teaching

Arrows denote factor→indicator standardized loadings based on the final CFA results. For readability, only a subset of interfactor covariances is displayed.

Discriminant Validity: Fornell–Larcker Criterion and HTMT

Discriminant validity was examined using two complementary approaches. First, the Fornell–Larcker criterion indicated that the square root of the AVE (√AVE) on the diagonal of the latent correlation matrix exceeded the inter-construct correlations, evidencing empirical distinction among dimensions (Fornell & Larcker, 1981). Second, the heterotrait–monotrait ratio (HTMT) fell below the conservative threshold of .85 (or

.90 for theoretically closely related constructs), further confirming construct separation at the level of inter-factor associations (Henseler et al., 2015). Full results are reported in Table 5 and 6.

Table 5. Discriminant Validity of Latent Correlations with √AVE on the Diagonal (Fornell–Larcker Criterion)

Parameters	PP	EFK	DK	PM	ERB	KK
PP	0.75					
EFK	0.62	0.81				
DK	0.54	0.58	0.69			
PM	0.68	0.71	0.65	0.79		
ERB	0.59	0.64	0.61	0.73	0.73	
KK	0.52	0.57	0.69	0.66	0.61	0.77

Table 6. Discriminant Validity of HTMT Matrix

Parameters	PP	EFK	DK	PM	ERB	KK
PP	-					
EFK	0.74	-				
DK	0.68	0.71	-			
PM	0.82	0.84	0.79	-		
ERB	0.73	0.78	0.76	0.87*	-	
KK	0.67	0.71	0.84	0.81	0.76	-

Asterisks indicate pairs with HTMT above .85 but below .90; interpretation is supported by theoretical justification.

Measurement Invariance (MI) Across Groups

MI was tested sequentially, configural (similar structure), metric (equality of loadings), and scalar (equality of intercepts), with stage-wise decisions based on the thresholds ΔCFI ≤ 0.010 and ΔRMSEA ≤ 0.015, which are more stable than the $\Delta\chi^2$ test (Chen, 2007; Cheung & Rensvold, 2002). A summary of MI results is provided in Table 6, and the procedural flow is depicted in Figure 2.

Figure 2. Measurement Invariance (MI) Workflow and Group-Specific Decision Panels. *Note.* If scalar invariance is not achieved, apply partial scalar invariance before conducting latent mean comparisons.

Gender. The configural model demonstrated adequate fit (e.g., $\chi^2/df \approx 2.10$; CFI ≈ .955; TLI ≈ .946; RMSEA ≈ .045). Constraining to the metric level yielded ΔCFI ≈ .002 and ΔRMSEA ≈ .001 (Pass), and further constraining to the scalar level yielded ΔCFI ≈ 0.005 and ΔRMSEA ≈ 0.002 (Pass). Implication: the model achieved scalar invariance; comparisons of latent means across gender are defensible (Little, 2013; Vandenberg & Lance, 2000).

Years of service (≤ 5 vs. > 5 years). The configural model fit was acceptable ($\chi^2/df \approx 2.30$; CFI = 0.952; TLI = .943; RMSEA = 0.047). The metric level passed (ΔCFI = .009; ΔRMSEA = 0.012), whereas the scalar level did not pass (ΔCFI = 0.010; ΔRMSEA = 0.017). Implication: comparisons of latent coefficients/correlations are appropriate (metric equivalence), whereas comparisons of latent means require partial scalar invariance by freeing intercepts of indicators identified as most problematic based on MI/LM evidence and theoretical rationale (Brown, 2015; Putnick & Bornstein, 2016). After freeing parameters, re-verify that changes in fit remain within ΔCFI ≤ 0.010 and ΔRMSEA ≤ 0.015; if satisfied, latent-mean comparisons may proceed with caution.

Certification (yes vs. no). The configural model was adequate (e.g., CFI ≈ 0.957; TLI ≈ 0.949; RMSEA ≈ .044); metric passed (ΔCFI ≈ .004; ΔRMSEA ≈ .002); scalar passed (ΔCFI ≈ 0.002; ΔRMSEA ≈ 0.003). Implication: the model achieved scalar invariance; latent-mean comparisons across certification groups are defensible (Little, 2013; Vandenberg & Lance, 2000).

Accordingly, the results in Table 6, which show scalar invariance for gender and certification and metric

invariance for years of service, are clearly reflected in Figure 2, enabling readers to trace the decision rationale and the analytical consequences (e.g., when latent-mean comparisons are defensible or when a partial scalar adjustment is required).

Table 7. Measurement Invariance (absolute Δ + group sizes)

Group	N1	N0	Level	χ^2/df	CFI	TLI	RMSEA	ΔCFI	ΔRMSEA	Decision
Gender (Male vs. Female)	9	21	Configural	2.10	0.955	0.946	0.045	—	—	Adequate fit
	9	21	Metric	—	—	—	—	0.002	0.001	Pass
	9	21	Scalar	—	—	—	—	0.005	0.002	Pass
Years of service (≤ 5 vs. >	11	19	Configural	2.30	0.952	0.943	0.047	—	—	Adequate fit
5 years)	11	19	Metric	—	—	—	—	0.009	0.012	Pass
	11	19	Scalar	—	—	—	—	0.010	0.017	Fail (scalar)
Certification (Yes vs. No)	18	12	Configural	2.05	0.957	0.949	0.044	—	—	Adequate fit
	18	12	Metric	—	—	—	—	0.004	0.002	Pass
	18	12	Scalar	—	—	—	—	0.002	0.003	Pass

Notes. ΔCFI and ΔRMSEA are absolute changes relative to the preceding model. Invariance decisions follow ΔCFI ≤ .010 and ΔRMSEA ≤ .015. N1/N0 are group sizes: Gender (male = 9; female = 21), Years of service (≤ 5 years = 11; > 5 years = 19), Certification (yes = 18; no = 12).

Criterion-Related Validity (Known-Groups)

Known-groups and theoretical implications. As an initial check of criterion-related validity, total/composite scores were compared between certified (n = 18) and non-certified (n = 12) teachers. The mean difference was practically significant (Cohen's $d \approx 0.63$; $p \approx 0.029$), supporting the hypothesis that certification is associated with higher readiness, consistent with literature positioning teacher readiness and efficacy as prerequisites for the effective implementation of CRT practices (Gay, 2018; Hammond, 2015). The finding is also consistent with the Funds of Knowledge (FoK) and psychological safety frameworks, in which teacher readiness mediates the integration of local knowledge with science and the cultivation of a psychologically safe classroom climate (Llopart & Esteban-Guitart, 2018; Manasia et al., 2020). Nonetheless, the causal relationship between certification and these indicators warrants further testing through longitudinal/experimental designs. Assumptions of variance equality were examined; Welch's test was used when appropriate, and 95% confidence intervals (CI) for d/Hedges' g were reported. Given that scalar MI was achieved for certification, latent mean comparisons across groups are defensible and consistent with composite-score results.

Sensitivity Analyses and Robustness Checks

Several checks were conducted to assess the robustness of the findings. First, a CFA model without additional modification indices produced only minor changes in fit and did not alter substantive conclusions, aligning with recommendations to prioritize theoretical rationale (Brown, 2015). Second, alternative estimation (e.g., standard ML) on a data subset with stronger normality assumptions yielded consistent patterns of fit and loadings. Third, alternative reliability (ω) supported the α-based findings, mitigating concerns about α's limitations in multidimensional models (Mcneish, 2017). Fourth, an examination of local dependence among similarly worded indicators did not reveal residual correlations that would compromise construct interpretation.

Comparison with the Literature, Implications, and Limitations

Comparison with the literature. High I-CVI/S-CVI values are consistent with content validity guidelines (Lynn, 1986; Polit & Beck, 2006), whereas satisfactory internal consistency (α, ω) and convergent–discriminant evidence (CR/AVE, Fornell–Larcker, HTMT) align with modern multitrait–multimethod recommendations (Fornell & Larcker, 1981; Hair et al., 2010; Henseler et al., 2015). Achieving scalar invariance for gender and certification advances the literature by providing an equitable instrument for latent mean comparisons across groups, an aspect often missing in studies within indigenous contexts. The absence of full scalar invariance for years of service, despite metric invariance, echoes methodological cautions that higher-level invariance is not always attainable and that partial scalar solutions are acceptable (Cheung & Rensvold, 2002; Putnick & Bornstein, 2016).

Practical implications. First, given adequate validity and reliability, the instrument can be used to map needs for in-service training more precisely, for example, strengthening CRT efficacy, enhancing institutional support for community collaboration, or developing contextualized science materials. Second, scalar invariance for certification permits defensible

latent mean comparisons to appraise certification effects on CRT readiness. Third, for years of service, comparisons should focus on latent coefficients/correlations (given metric equivalence) or proceed via partial scalar invariance with transparent reporting, thereby ensuring evidence-based decision-making (Little, 2013; Vandenberg & Lance, 2000).

Limitations. The pilot sample size (N ≈ 30) was adequate for preliminary reliability and item statistics but limited for more granular inference. Nevertheless, the layered strategy—CVI, face validity, pilot testing, CFA, and MI, enhanced the credibility of the findings. Second, the known-groups criterion indicator (certification) should not be interpreted as causal; although a medium effect size strengthens nomological validity, replication in longitudinal or experimental designs is needed. Third, the highly specific local context (borderland indigenous communities) warrants caution in generalization; however, the transparent validation framework facilitates adaptation to other indigenous communities in Indonesia, with rigorous cultural adaptation procedures (American Educational Research Association, American Psychological Association, 2014).

Summary of Hypothesis Tests

H1 (adequate content validity): Supported, all items achieved I-CVI ≥ 0.78 and S-CVI/Ave ≥ 0.90 (Lynn, 1986; Polit & Beck, 2006).

H2 (adequate internal consistency and sound item statistics): Supported, α = 0.86; ω = 0.87; r_it ≥ .30; floor/ceiling < 15% (Hair et al., 2010; Mcneish, 2017).

H3 (construct validity): Supported, CFA indicated acceptable fit; CR ≥ 0.70; AVE ≥ 0.50; discriminant validity satisfied (Brown, 2015; Fornell & Larcker, 1981; Henseler et al., 2015; Kline, 2016).

H4 (measurement invariance): Largely supported, scalar invariance for gender and certification; metric invariance for years of service (partial scalar required for latent mean comparisons) (Chen, 2007; Cheung & Rensvold, 2002; Putnick & Bornstein, 2016).

H5 (initial criterion-related validity): Supported, total-score differences between certified and non-certified teachers showed a medium effect size, consistent with CRT theory (Gay, 2018; Hammond, 2015).

Discussion

Interpreting the Main Findings in CRT Theory

This study successfully developed and validated a teacher readiness instrument for Culturally Responsive Teaching (CRT) that is psychometrically adequate for primary schools in the Indonesia–Papua New Guinea border region. High content validity (I-CVI ≥ 0.78; S-CVI/Ave ≥ 0.90) confirms that the indicators represent the CRT construct domain comprehensively, aligning with theoretical frameworks that treat culture as both an epistemic and pedagogical resource (Gay, 2018).

The multidimensional structure confirmed by CFA-CRT pedagogical knowledge, CRT efficacy, school contextual support, culturally responsive planning & materials, culturally responsive assessment, and community collaboration is consistent with Hammond (2015) comprehensive model of essential components in CRT implementation. These findings reinforce the view that CRT teacher readiness is not a unidimensional construct but a complex configuration of knowledge, beliefs, skills, and contextual support that interact with one another. The knowledge/skills dimension encompasses understanding cultural contexts and pedagogical strategies for designing inclusive curricula and assessments (Hu et al., 2021); the beliefs/attitudes dimension concerns how efficacy and cultural identity awareness shape classroom practice (e.g., evidence from the Identity Project intervention) (Pevec-Zimmer et al., 2024); while contextual support emphasizes organizational/system readiness that enables teachers' preparedness in practice (Wang et al., 2020). Accordingly, an exclusive focus on cognitive–technical aspects (knowledge/skills) risks overlooking emotional–psychological dimensions that are also crucial for truly inclusive classrooms (Manasia et al., 2020).

Adequate internal reliability (α = 0.86; ω = 0.87) indicates dependable measurement consistency, while convergent and discriminant validity evidence suggests that the instrument's dimensions are interrelated yet empirically distinct. This is important for identifying teacher-specific readiness profiles that can inform more targeted professional development designs.

The Significance of Measurement Invariance for Indigenous Contexts

Achieving scalar invariance for gender and certification is a notable methodological contribution because it allows for defensible latent mean comparisons across those groups. This finding addresses a common methodological limitation in instrument studies conducted in indigenous contexts, where measurement equivalence is seldom examined systematically (Putnick & Bornstein, 2016).

The failure to reach full scalar invariance for years of service (≤ 5 vs. > 5 years), despite metric invariance, reflects the complexity of professional dynamics in border regions. It suggests that teachers with different tenure lengths may interpret aspects of CRT readiness differently, potentially due to variations in community engagement experience and evolving understandings of culturally responsive practice over time.

Implementing partial scalar invariance for years of service, by freeing intercepts of the most problematic indicators, offers a practical solution enabling limited yet meaningful comparisons (Brown, 2015). Transparent reporting of this procedure is essential to ensure accurate interpretation and replicability.

Criterion Validity and Practical Relevance

Initial criterion-related evidence, differences between certified and non-certified teachers ($d \approx 0.63$), supports the argument that the instrument is sensitive to theoretically expected distinctions. This finding is consistent with literature indicating that teacher certification, while imperfect, is generally associated with higher pedagogical competence (Darling-Hammond et al., 2020).

However, cautious interpretation is warranted. The observed differences may reflect not only the impact of certification per se but also related factors such as training experiences, access to professional resources, or personal characteristics that influence motivation to pursue certification. Longitudinal or experimental research is needed to establish a more definitive causal relationship.

Contribution to CRT Literature in Non-Western Contexts

This study fills an important gap by providing empirical evidence from a non-Western indigenous context. Most existing CRT instruments were developed with urban or multicultural populations in high-income countries, with limited representation of indigenous communities (Hernandez, 2022; Paris, 2012). Validating the instrument in the Indonesia, Papua New Guinea border context demonstrates that the CRT construct is relevant and can be operationalized validly beyond its original development settings.

Cultural adaptation in this study, lexical adjustments, locally contextualized examples, and involvement of local teachers, reflects practices aligned with indigenous research methodologies, which emphasize community participation and authentic representation to ensure relevance and respect for indigenous contexts (Ryder et al., 2020; Snow et al., 2016). Practically, cultural adaptation was implemented through lexical adjustments to maintain intelligibility and linguistic proximity, integration of contextual examples resonant with community experience, and the involvement of local teachers in face validity as a form of co-production of knowledge. These adaptations model rigorous cultural adaptation practices that can be replicated in other indigenous communities in Indonesia. This approach accords with indigenous methodology principles, participation and authentic representation, while acknowledging potential challenges (e.g., resistance from external researchers unfamiliar with methodological nuances), thereby underscoring the need for ongoing dialogue and methodological literacy (Cummings, 2020; Datta, 2018; Windchief & Cummins, 2021).

Implications for Teacher Professional Development

The instrument's multidimensional structure provides a diagnostic framework to inform more specific and responsive professional development. For example, teachers with high CRT pedagogical knowledge but low community collaboration scores could receive interventions focused on building school–community partnerships.

Scalar invariance for certification allows education providers to evaluate certification programs' effectiveness in enhancing CRT readiness using defensible latent mean comparisons—vital for accountability and continuous improvement.

In resource-constrained border settings, the instrument can facilitate more efficient allocation of interventions by identifying the dimensions most in need of strengthening. Such an evidence-based approach can increase the impact of limited professional development resources.

Limitations and Their Implications

Several limitations should be acknowledged. First, the pilot sample ($N \approx 30$) constrains the stability of parameter estimates and the generalizability of preliminary findings. Although the layered validation strategy increases credibility, replication with larger samples is required. Second, the cross-sectional design does not permit inferences about temporal stability or sensitivity to change in teacher readiness. For program monitoring and evaluation, evidence on test–retest reliability and responsiveness is needed. Third, criterion validity was limited to a simple known-groups comparison. Predictive validity linking instrument scores to student learning outcomes or classroom practice quality would substantively strengthen the validity argument. Fourth, the highly specific local context (Papua borderland indigenous communities) warrants caution in generalization. While the transparent validation framework facilitates adaptation, each application in a different context requires independent psychometric verification.

Directions for Future Research

Several avenues can extend these contributions. First, longitudinal studies tracking changes in teacher readiness and its relation to student outcomes would strengthen predictive validity and practical utility. Second, research exploring mediators and moderators between teacher readiness and CRT implementation effectiveness can clarify underlying causal mechanisms;

factors such as administrative support, community characteristics, and school resources may play crucial roles. Third, adapting and validating the instrument in other indigenous contexts in Indonesia can test generalizability and support the development of more representative norms; a meta-analytic approach across contexts could identify universal versus culture-specific elements of CRT readiness. Fourth, developing a short form for routine monitoring could enhance feasibility at scale while retaining essential psychometric properties.

Policy and Practice Implications

The findings have direct implications for education policy in border regions and areas with indigenous populations. First, the instrument can be integrated into teacher competency appraisal systems to ensure that cultural responsiveness receives adequate attention in performance evaluation. Second, readiness profiles can inform allocation of scholarships or incentives for professional development, prioritizing teachers with high motivation but improvable readiness. Third, aggregated data from instrument implementation can inform teacher education curricula, ensuring that programs prepare candidates with sufficient CRT competencies from the outset of their careers.

Conclusions

This study successfully developed and validated a teacher readiness instrument for Culturally Responsive Teaching (CRT) that is psychometrically adequate for primary schools in the Indonesia–Papua New Guinea border region. Through a rigorous, layered validation approach—covering content validity, face validity, pilot testing, confirmatory factor analysis, and measurement invariance—the study provides comprehensive empirical evidence on the measurement quality of the instrument. All research questions were addressed with results that largely support the hypotheses. Content validity met stringent standards, with all items achieving I-CVI ≥ 0.78 and S-CVI/Ave ≥ 0.90, confirming adequate coverage of the construct domain. Internal consistency was good (Cronbach's α = .86; McDonald's ω = 0.87), and item statistics indicated healthy distribution and discrimination. The theoretically specified multidimensional factor structure, CRT pedagogical knowledge, CRT efficacy, school contextual support, culturally responsive planning & materials, culturally responsive assessment, and community collaboration, was confirmed via CFA with acceptable fit indices. Convergent and discriminant validity were satisfied according to standard criteria (factor loadings, AVE, CR, Fornell–Larcker, and HTMT), indicating that dimensions are interrelated yet empirically distinct. The attainment of measurement invariance represents a notable methodological contribution. Scalar invariance was achieved for gender and certification, enabling defensible latent mean comparisons across these groups. Although years of service achieved only metric invariance, a partial scalar invariance solution offers an acceptable basis for limited comparisons with transparent reporting.

Acknowledgements

The authors gratefully acknowledge the financial support of the Ministry of Higher Education, Science, and Technology through the Fundamental Research Scheme, under Master Contract No. 065/C3/DT.05.00/PL/2025 dated 28 May 2025. This support facilitated the entire research process, from design to manuscript preparation.

Author Contribution

This article was written by three authors, namely A. A., A. K., and R. R. All authors contributed to each stage of the research carried out.

Funding

This research fund by Ministry of Higher Education, Science, and Technology through the Fundamental Research Scheme, under Master Contract No. 065/C3/DT.05.00/PL/2025.

Conflicts of Interest

All authors declare no conflict interest in this article.

Reference

Allison-Burbank, J. D., Conn, A., & Vandever, D. (2023). Interpreting Diné Epistemologies and Decolonization to Improve Language and Literacy Instruction for Diné Children. *Language Speech and Hearing Services in Schools*, *54*(3), 707–715. https://doi.org/10.1044/2023_lshss-22-00147

Andrew, & Johnson. (2022). Culturally Responsive Teaching in Higher Education. *International Journal of Equity and Social Justice in Higher Education*. https://doi.org/10.56816/2771-1803.1008

Arlianto, A., Prahoro, A. P., & Oktavia, A. (2024). Intercultural competence measurement tools for Indonesian students: Adaptation, testing construct validity, and measurement invariance with the MIMIC model. *Jurnal Psikologi UNDIP*, *23*(1), 1–24. https://doi.org/10.14710/jp.23.1.1-24

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). The Guilford Press.

Chen, F. F. (2007). Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling: A*

Multidisciplinary Journal, *9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Cummings, J. (2020). Review for "Trajectories of peer victimization in elementary school children: Associations with changes in internalizing, externalizing, social competence, and school climate. *Wiley*. https://doi.org/10.1002/jcop.22365/v1/review1

Darling-Hammond, L., Flook, L., Cook-Harvey, C., Barron, B., & Osher, D. (2020). Implications for educational practice of the science of learning and development. *Applied Developmental Science*, *24*(2), 97–140. https://doi.org/10.1080/10888691.2018.1537791

Datta, R. (2018). Traditional storytelling: an effective Indigenous research methodology and its implications for environmental research. *AlterNative*, *14*(1), 35–44. https://doi.org/10.1177/1177180117741351

Fornell, C., & Larcker, D. F. (1981). Evaluating Structural Equation Models with Unobservable Variables and Measurement Error. *Journal of Marketing Research*, *18*(1), 39–50. https://doi.org/10.1177/002224378101800104

Gay, G. (2018). *Culturally responsive teaching: Theory, research, and practice*. New York: Teachers College Press.

Greenfader, C. M. (2022). Latinx Family Engagement in Early Elementary School: A National Study. In *AERA*. https://doi.org/10.3102/ip.22.1891383

Hair, J. F., Black, W. J., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis*. Englewood Cliff, Prentice Hall.

Hammond, Z. (2015). *Culturally Responsive Teaching and The Brain*. SAGE Publications.

Henseler, J., Ringle, C. M., & 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. https://doi.org/10.1007/s11747-014-0403-8

Hernandez, A. (2022). Closing the Achievement Gap in the Classroom Through Culturally Relevant Pedagogy. *Journal of Education and Learning*, *11*(2), 1. https://doi.org/10.5539/jel.v11n2p1

Hu, X., Xu, Z., Neshyba, M. V, Geng, Z., & Turner, R. K. (2021). A multi-dimensional model: implications for preparing pre-service teachers for culturally responsive teaching. *Asia-Pacific Journal of Teacher Education*, *49*(3), 282–299. https://doi.org/10.1080/1359866X.2020.1753169

Hung, H.-F., Yen, C., Yao, T., & Lin, S. (2023). Exploring How Place-Based Education Indigenous Curriculum Influence Students. In *Learning Motivation in Science* (pp. 191–215). Springer International Publishing. https://doi.org/10.1007/978-3-031-30506-1_11

Joseph, D. H., Keene, C., Castagno, A. E., Dass, P. M., & Macias, C. (2024). Methodological Complexity: A Both/and Approach to Address Tool Validity and Reliability for Assessment of Cultural Responsiveness in Indigenous Serving Schools. *AERA Open*, *10*. https://doi.org/10.1177/23328584241232958

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

List, A., Campos Oaxaca, G. S., Du, H., Lee, H. Y., & Lyu, B. (2024). Critical culturalized comprehension: Exploring culture as learners thinking about texts. *Educational Psychologist*, *59*(1), 1–19. https://doi.org/10.1080/00461520.2023.2266028

Little, T. D. (2013). *Longitudinal structural equation modeling*. The Guilford Press.

Llopart, M., & Esteban-Guitart, M. (2018). Funds of knowledge in 21st century societies: inclusive educational practices for under-represented students. A literature review. *Journal of Curriculum Studies*, *50*(2), 145–161. https://doi.org/10.1080/00220272.2016.1247913

Lynn, M. R. (1986). Determination and Quantification Of Content Validity. *Nursing Research*, *35*(6). Retrieved from https://journals.lww.com/nursingresearchonline/fulltext/1986/11000/determination_and_quantification_of_content.17.aspx

Manasia, L., Ianos, M. G., & Chicioreanu, T. D. (2020). Pre-service teacher preparedness for fostering education for sustainable development: An empirical analysis of central dimensions of teaching readiness. *Sustainability (Switzerland)*, *12*(1). https://doi.org/10.3390/SU12010166

Mcneish, D. (2017). Psychological methods: Thanks coefficient Alpha, we'll take it from here. *Psychological Methods*, 1–22. https://doi.org/10.1037/met0000144

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(ue 9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Panggabean, M. S., & Himawan, K. K. (2016). The Development of Indonesian Teacher Competence Questionnaire. *Journal of Educational, Health and Community Psychology*, *5*(2), 1–15. https://doi.org/10.12928/JEHCP.V5I2.5134

Paris, D. (2012). Culturally Sustaining Pedagogy. *Educational Researcher*, *41*(3), 93–97. https://doi.org/10.3102/0013189x12441244

Pevec-Zimmer, S., Juang, L. P., & Schachner, M. K. (2024). Promoting awareness and self-efficacy for culturally responsive teaching of pre-service teachers through the identity project–a mixed methods study. *Identity: International Journal of Theory and Research*, *24*(4), 288–306. https://doi.org/10.1080/15283488.2024.2344086

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? critique and recommendations. *Research in Nursing & Health*, *29*(5), 489–497. https://doi.org/10.1002/nur.20147

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Journal Developmental Review*, *41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

Razfar, A., & Nasir, A. (2019). Repositioning English Learners' Funds of Knowledge for Scientific Practices. *Theory Into Practice*, *58*(3), 226–235. https://doi.org/10.1080/00405841.2019.1599231

Ryder, C., Mackean, T., Coombs, J., Williams, H., Hunter, K., Holland, A. J. A., & Ivers, R. (2020). Indigenous Research Methodology -- Weaving a Research Interface. *International Journal of Social Research Methodology*, *23*(3), 255–267. https://doi.org/10.1080/13645579.2019.1669923

Snow, K. C., Hays, D. G., Caliwagan, G., Ford, D. J., Mariotti, D., Mwendwa, J. M., & Scott, W. (2016). Guiding principles for indigenous research practices. *Action Research*, *14*(4). https://doi.org/10.1177/1476750315622542

Sotero, M. C., Alves, Â. G. C., Arandas, J. K. G., & Medeiros, M. F. T. (2020). Local and scientific knowledge in the school context: Characterization and content of published works. *Journal of Ethnobiology and Ethnomedicine*, *16*(1). https://doi.org/10.1186/s13002-020-00373-5

Vandenberg, R. J., & Lance, C. E. (2000). A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Methods*, *3*(1), 4–70. https://doi.org/10.1177/109442810031002

Villegas, A. M., & Lucas, T. (2007). *The Culturally Responsive Teacher*. Educational Leadership.

Wang, T., Wang, T., Olivier, D. F., & Chen, P. (2020). Creating individual and organizational readiness for change: conceptualization of system readiness for change in school education. *International Journal of Leadership in Education*, 1–25. https://doi.org/10.1080/13603124.2020.1818131

Windchief, S., & Cummins, J. D. (2021). Considering Indigenous Research Methodologies: Bicultural Accountability and the Protection of Community Held Knowledge. *Qualitative Inquiry*. https://doi.org/10.1177/10778004211021803

Yektiningtyas, W., Wompere, R. N. N., Kobepa, N., & Sunarsih, T. A. (2023). Engaging students to write procedure texts through the culturally-relevant activity of bark painting. *JOALL (Journal of Applied Linguistics and Literature)*, *8*(1), 41–58. https://doi.org/10.33369/joall.v8i1.22577