Academia Open

Vol. 10 No. 2 (2025): December DOI: 10.21070/acopen.10.2025.12776

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Academia Open

Vol. 10 No. 2 (2025): December DOI: 10.21070/acopen.10.2025.12776

Table Of Contents

Journal Cover 1
Author[s] Statement 3
Editorial Team 4
Article information 5
Check this article update (crossmark) 5
Check this article impact 5
Cite this article 5
Title page 6
Article Title 6
Author information 6
Abstract 6
Article content 6

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Academia Open

Vol. 10 No. 2 (2025): December DOI: 10.21070/acopen.10.2025.12776

Originality Statement

The author[s] declare that this article is their own work and to the best of their knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the published of any other published materials, except where due acknowledgement is made in the article. Any contribution made to the research by others, with whom author[s] have work, is explicitly acknowledged in the article.

Conflict of Interest Statement

The author[s] declare that this article was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright Statement

Academia Open

Vol. 10 No. 2 (2025): December DOI: 10.21070/acopen.10.2025.12776

EDITORIAL TEAM

Academia Open

Vol. 10 No. 2 (2025): December DOI: 10.21070/acopen.10.2025.12776

Article information

Check this article update (crossmark)

Check this article impact (*)

Save this article to Mendeley

(*) Time for indexing process is various, depends on indexing database platform

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Academia Open

Vol. 10 No. 2 (2025): December
DOI: 10.21070/acopen.10.2025.12776

Validated Culturally Responsive Science Assessment Using Integrated Content and Construct Analysis

Aisyah Ali, aisyahali@fkip.uncen.ac.id (1)

Primary teacher education, Universitas Cenderawasih, Jayapura, Indonesia

Singgih Bektiarso, singgih.fkip@unej.ac.id (0)

Doctoral Program of Science Education, Faculty of Teacher Training and Education, Universitas Negeri Jember, Jember, Indonesia

Auldry Fransje Walukow, auldrywalukow@fkip.uncen.ac.id (0)

Doctoral Study Programs in Science Education, Universitas Jember, Jember, Indonesia

Erlia Narulita, erlia.fkip@unej.ac.id (0)

Physics Education, Universitas Cenderawasih, Jayapura, Indonesia

Akhmad Kadir, akhmad.kadir@fisip.uncen.ac.id (0)

Department of Anthropology, Universitas Cenderawasih, Jayapura, Indonesia

(1) Corresponding author

Abstract

General Background: The alignment of science assessment with students' socio-cultural contexts is essential to ensure fairness and meaningful measurement of learning outcomes. **Specific Background:** However, most contextualized assessments in science education emphasize content validity without empirically confirming their construct structure, limiting their interpretive strength. **Knowledge Gap:** There remains a lack of studies that integrate content, empirical, and construct validity evidence in culturally responsive instruments, particularly those designed in parallel pre–post forms. **Aims:** This study aimed to develop and validate an ethnoscience-based pre–post instrument by linking Aiken's Content Validity Ratio (CVR, 4-point scale) with Confirmatory Factor Analysis (CFA, CR/AVE). **Results:** Findings from five expert reviews showed 22 of 40 items exceeded the conservative threshold (Aiken's V ≥ 0.80; CVR = 1.00). Field trials (N = 50) demonstrated moderate difficulty and positive discrimination, while CFA confirmed a three-factor structure with good fit (χ^2 = 34.203, df = 24, p = 0.083; CFI = 0.94; TLI = 0.92; RMSEA = 0.065). Composite reliability ranged from 0.718–0.797, and AVE was adequate for two factors (0.506; 0.568) and marginal for one (0.459). **Novelty:** The study presents a transparent "content–empirical–construct" decision trail rarely reported in ethnoscience assessment. **Implications:** This integrative validation framework demonstrates that cultural responsiveness and psychometric rigor can coexist, guiding fair and contextual science learning evaluations.

Highlights:

- Integrates Aiken–CVR and CFA for comprehensive validity evidence.
- Confirms three-factor model with strong reliability and moderate AVE.
- Demonstrates synergy between cultural relevance and measurement rigor.

Keywords: Content Validity, Confirmatory Factor Analysis, Ethnoscience, Culturally Responsive Assessment, Psychometric Validation

Academia Open

Vol. 10 No. 2 (2025): December
DOI: 10.21070/acopen.10.2025.12776

Published date: 2025-10-20

Introduction

The development of science learning outcome assessments that align with the socio-cultural context of students is a prerequisite for fair, relevant, and effective learning. Within the framework of culturally responsive pedagogy, test items should not merely inventory concept memorization but rather stimulate scientific reasoning through stimuli close to students' life experiences; thus, scores better represent the targeted conceptual competencies and cognitive processes, rather than merely linguistic skills or familiarity with test culture.[1] These needs have driven the design of an ethnoscience-based instrument with two parallel forms (pre–post) aligned with learning objectives and cognitive taxonomy, as well as a layered validation pathway from content evidence, item empirical evidence, to construct evidence at the latent measurement model level. This layered approach is oriented to meet the expectations of reputable publications: a clear theoretical foundation, rigorous methodology, and a tangible contribution to assessment practices that respect and empower the local cultural context.[2], [3]

At the stage of content validity, a panel of five experts assessed each item on five core criteria—clarity, relevance, domain representation, alignment with objectives, and editorial bias—using a four-point scale for calculating Aiken's V and a binary essential/not essential decision for the Content Validity Ratio (CVR). This combination captures two dimensions of content evidence coverage: graded assessment intensity (Aiken) and essentiality consensus (CVR). The choice of a four-point scale was maintained to reinforce expert decisions and avoid midpoint ambiguity, as recommended in the methodological literature on content validity in the early stages of instrument development.[4], [5], [6]. To maintain conservatism, the summary per item uses V_min (the lowest V value among the criteria) as the basis for the initial KEEP/REVIEW decision, while CVR is reported as supporting evidence, considering that critical values for a small panel (N≈5) are relatively strict.[7] This approach aligns with the practice of developing ethnoscience-based educational instruments that emphasize content accountability and cultural sensitivity simultaneously.[3]

After expert review, a limited field trial was conducted to estimate item parameters—difficulty level, discrimination, and distractor functioning as an empirical "safety net" to verify content-based decisions. Revisions focused on three axes: (i) clarifying wording to reduce linguistic bias (e.g., lexical absolutism and double negation), (ii) aligning stimulus–indicator to make the required cognitive evidence clearly evident, and (iii) standardizing option/distractor patterns to minimize testwiseness, while maintaining balance in domain coverage and cognitive level distribution so that pre–post parallelism is preserved.[8] The entire process is supported by design trail documentation—grids, expert evaluation sheets, and decision logs—to ensure procedure traceability and the possibility of replication.[1] Construct confirmation was then carried

Vol. 10 No. 2 (2025): December
DOI: 10.21070/acopen.10.2025.12776

out through Confirmatory Factor Analysis (CFA) on a three-factor model derived from the domain × cognitive process blueprint. Reporting included model fit indices (χ^2, df, p-value; CFI; TLI; RMSEA with 90% CI; and SRMR), standardized factor loadings (λ), composite reliability (CR), and convergent validity (AVE). Within this framework, readers not only assess 'model fit' but also the strength of indicator loadings and the proportion of variance explained by the construct—two prerequisites essential for valid and defensible score interpretation in pre-post evaluation. Such reporting practices are consistent with scale development standards that require alignment between conceptual definitions and empirical evidence.[5]

From a state-of-the-art perspective, many reports on the development of contextual instruments stop at content validity (e.g., Aiken's V) and basic reliability; relatively few explicitly link KEEP/REVIEW decisions to CFA findings (loadings, CR, AVE) after trying them out. On the other hand, the literature also emphasizes complementary quantitative content evidence such as the Content Validity Index (CVI) and expert agreement reliability, for example through multi-rater kappa, to strengthen the interpretation of design decisions.[7], [9]. Thus, there is a gap between the methodological ideal, namely the chain of evidence from content to construct, and actual reporting practices, which often break at the content phase. This manuscript addresses this gap by structuring the decision trail from content evidence → item-level empirical evidence → construct evidence, while emphasizing the importance of pre–post form parallelism (alignment of goals, domains, and cognitive demands) so that score differences are more likely to reflect instructional gain rather than test artifacts.[1]

Based on the methodological needs outlined above, the research questions are formulated as follows. First, how can it be ensured that the two parallel forms (pre and post) derived from the domain × Bloom blueprint are truly equivalent in terms of objectives, content, and cognitive demands so that the pre–post score comparison is fair? Second, to what extent is content validity satisfied based on Aiken's V (four-point scale) and essentiality consensus (CVR) when five cross-disciplinary experts review the items against core criteria? Third, how does the empirical performance of the items in a limited trial look in terms of difficulty, discrimination, and distractor function, and what are the implications for revision decisions while maintaining domain coverage and distribution of cognitive levels? Fourth, is the hypothesized construct structure confirmed through CFA, and how do the CR/AVE profiles for each factor guide the refinement of subsequent indicators?

Methodologically, this study illustrates the recommended development pathway: (optional) EFA → CFA with reporting of model fit indices and CR/AVE; beyond that, CFA is treated not merely as a formal step but as an epistemic mechanism to test the claim that culturally grounded indicators indeed load on the target construct. To enhance external validity and fairness of interpretation in the future, we direct further research toward measurement invariance testing (configural–metric–scalar), anchor-item-based form equating, and sample expansion to improve parameter stability.[10], [11]. Thus, the integration of cultural relevance and the rigor of measurement offered in this study is expected to contribute to a science assessment framework

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Vol. 10 No. 2 (2025): December
DOI: 10.21070/acopen.10.2025.12776

that is fairer, more contextual, and accountable, in line with the agenda of improving science learning quality in culturally diverse environments.

Method

This study uses a layered development–validation design that integrates content evidence (through expert assessment based on Aiken's V and Lawshe's CVR), empirical evidence (item analysis in a limited field trial), and construct evidence (confirmation of the measurement model through CFA). This approach aligns with methodological recommendations that instrument validity should be supported by a chain of evidence from content to construct, rather than relying on a single indicator. In the content stage, a four-point scale was used to make expert evaluations more decisive and avoid midpoint ambiguity. CVR is used as a measure of "essential/not essential" consensus to complement information on rating intensity.[6] The model construction stage uses CFA to examine the fit of the latent structure, factor loadings (λ), composite reliability (CR), and convergent validity (AVE) so that score interpretations can be justified at the construct level.[5]

The research involved five cross-disciplinary experts as an expert panel for content validity, in accordance with common practice in early-stage studies [5], and a trial sample of N = 50 students (demographic/curriculum details filled according to field data). Expert recruitment considers substantive expertise (related scientific domain), assessment methodology, and representation of local cultural context. Participation is voluntary with informed consent and protection of aggregated data confidentiality.

Table 1. Characteristics of the Expert Panel

Expert System	Field of Expertise	Length of Experience (years)	Role (Academician/Practitioner)	Test Development Experience (brief)
E1	Science Education	20	Academician	Bank coordination on; parallel pre–post; domain×Bloom alignment.
E2	Culture/Anthropology	12	Academics–Practitioners	Curation of ethnoscience contexts; verification of cultural sensitivity; co-design of stimuli.
E3	Language/Linguistics	8	Academician	Language review; editorial bias mitigation; adaptation of two languages/dialects.
E4	Assessment/Psychometrics	8	Academician	Design–validation; Aiken–CVR; item analysis; CFA, CR/AVE.
E5	Education evaluator/researcher & data analyst	15	Academician-Practitioner	Pre–post evaluation design; data analytics; audit decision log

Vol. 10 No. 2 (2025): December
DOI: 10.21070/acopen.10.2025.12776

1. Instrument Development

Items are developed from the domain blueprint × cognitive processes (based on Bloom's taxonomy), with culturally responsive principles: item stimuli and context are linked to local practices/artifacts to trigger meaningful scientific reasoning, not just fact recall. Two parallel forms (pre–post) are designed to align in objectives, content, and cognitive demands to support fair assessment of learning changes. Item drafts undergo review by content experts before entering Aiken–CVR quantitative evaluation.[5]

2. Content Validity Procedure: Aiken's V (4-Point Scale) and Lawshe's CVR

A panel of experts evaluated each item based on five criteria: (1) clarity, (2) relevance, (3) domain representation, (4) alignment with objectives, and (5) editorial bias (reverse) on a four-point scale (1 = very inappropriate; 4 = highly appropriate). According to the framework (Aiken, 1985), the content validity of each item is calculated using Aiken's V coefficient, a quantitative measure that aggregates the assessments of n experts to estimate the extent to which an item represents the construct being measured. In parallel, the CVR is calculated from the binary decision "essential/not essential" according to Lawshe.[6]. In parallel, CVR is calculated from a binary decision of "essential/not essential." If more than half of the panelists indicate that an item is important/essential, then the item has at least adequate content validity. The initial decision rule is to KEEP if Vmin ≥ 0.80V and REVIEW otherwise. CVR is reported for consensus transparency. As a control for content reporting, we also consider CVI practices (i-CVI/s-CVI) and expert agreement reliability (e.g., multi-rater kappa) as recommended in the literature. [9], However, the main focus remains on Aiken–CVR for consistency with the early development stages.

3. Field Trials and Item Analysis

The instrument was tested on a limited basis (N = 50) to obtain empirical evidence regarding difficulty level, discrimination, and distractor function. Item performance criteria refer to basic psychometric practices: difficulty distribution is within the moderate range; item–total (or point-biserial) correlations are positive; and distractors are chosen relatively more often by low-ability groups and less frequently by high-ability groups. Revisions focused on clarifying wording (avoiding lexical absolutism/double negatives), sharpening the linkage of indicator stimuli, and balancing option length structures to prevent testwiseness.

4. Confirmation Construct : CFA, CR, and AVE

Construct confirmation was carried out using Confirmatory Factor Analysis (CFA) to test the fit of the data with the three-factor measurement model derived from the domain × cognitive process blueprint. Estimation was performed using the maximum likelihood approach, and parameters were reported in standardized form (loading, SE, z, p) so that the strength of the indicators' loadings on the construct could be assessed transparently. [12], [13]. The model's suitability is comprehensively evaluated through a combination of commonly used indices, namely χ^2 (df, p), CFI, TLI, RMSEA with 90% CI, and SRMR. Interpretation relies on contemporary conventions: CFI/TLI ≥ 0.90 generally indicates adequate fit; a small

Vol. 10 No. 2 (2025): December
DOI: 10.21070/acopen.10.2025.12776

RMSEA with the upper bound of the 90% CI < 0.10 and a low SRMR support model suitability, noting that the p-value of χ^2 should be interpreted cautiously given its sensitivity to sample size.[13], [14], [15].

In addition to the global index, the evaluation of the internal structure emphasizes the loading of significant standardized factors (λ) that are within a moderate to high range, consistent with the conceptual definition of the construct. [12], [13]. To assess internal consistency at the construct level, Composite Reliability (CR) is reported; whereas convergent validity is summarized through the Average Variance Extracted (AVE). Operational criteria follow common scale development practice: CR ≥ 0.70 is interpreted as adequate construct reliability, and AVE ≥ 0.50 indicates that a sufficient proportion of indicator variance is explained by the construct (Fornell & Larcker, 1981). If a factor is found to have a marginal AVE but CR remains ≥ 0.70 and indicator loadings are significant, the findings are classified as adequate convergent validity but require indicator refinement in the next cycle.[12], [16].

Revision decisions do not rely on a single index alone, but on triangulation among model fit, loading profiles and their significance, and CR/AVE patterns. Inter-factor covariances are reported to show substantive relationships without causing conceptual redundancy; potential model changes are only considered if consistent with theory and supported by reasonable diagnostic indicators (e.g., modification indices).[13] All decisions regarding the retention, merging, or reduction of indicators are documented in the decision log and linked to the accompanying evidence. This practice maintains a methodological audit trail and facilitates replication. In line with early-stage content validity best practice recommendations, CFA results are also reintegrated into the content context: findings at the construct level (fit, λ, CR/AVE) are linked to design decisions at the item level (Aiken's V four-point scale for rating intensity, CVR for essentiality consensus, and—if used—multi-rater CVI/kappa for consensus reliability), so that the evidence path from content to construct is complete and accountable.[4], [13], [17]

5. Data Analysis Procedure

The entire content validity calculation (Aiken–CVR), item analysis, and initial decision summary were conducted on a structured worksheet (Aiken–CVR template; automatic calculation of Vmin, CVR, and revision flags). CFA analysis was performed using statistical software that supports structural equation modeling. Reproducibility was maintained through an analytic codebook and a decision log recording item/factor changes after result review.

6. Ethical Considerations

The ethics protocol includes written informed consent, confidentiality of individual data, and reporting results in aggregate. All procedures comply with institutional ethical guidelines and best practices in educational research. Participation is voluntary and can be withdrawn at any time without consequences.

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Vol. 10 No. 2 (2025): December
DOI: 10.21070/acopen.10.2025.12776

Results and Discussion

A. Summary of Content Validity (Aiken–CVR, Four-Point Scale)

Assessment by five experts produced a strong content validity pattern for most items. Referring to Aiken's V for the intensity of the graded ratings (scale 1–4) and the Content Validity Ratio (CVR) for essentiality consensus (binary), a conservative item-by-item summary uses the Vmin statistic, that is, the lowest Aiken value among the five criteria (clarity, relevance, domain representation, alignment with objectives, and reversed editorial bias), as is common in early-stage reviews.[4], [6]. Following the established decision rules, KEEP is granted when Vmin ≥ 0.80V and **CVR = 1.00; the rest are labeled REVIEW as refinement candidates. Based on computations from the analytic file, out of 40 items assessed, 22 were marked KEEP and 18 REVIEW. These findings indicate that the majority of items have met the expected intensity and consensus thresholds in the content phase, in line with the recommendation to use a four-point expert scale to avoid middle-choice ambiguity and reinforce panel decisions.[7].

Qualitatively, items that fall into the REVIEW category generally show one or two criteria with relatively lower V values (for example, in clarity or alignment of objectives), or CVR that has not reached full agreement among five experts. This is consistent with the nature of CVR, which tends to be strict on small panels, so it is recommended as supporting evidence, not the sole determinant, and is used to guide the gradual revision of items.[6]. All decisions (KEEP/REVIEW) are documented in the decision log to maintain traceability and facilitate replication [5].

B. Limited Trials and Item Analysis

A limited field trial (N = 50) was intended to examine the empirical consistency of items before construct confirmation. Descriptive results show that the difficulty distribution tends to be moderate, most discrimination indices are positive, and the distractors function as designed (relative frequency is higher in low-ability groups and decreases in high-ability groups). Methodologically, this evidence serves as a "safety net" for content-based decisions and provides feedback for improving phrasing (e.g., reducing double negatives/lexical absolutism), aligning stimulus–indicator, and balancing options to minimize testwiseness, a practice recommended to bridge the gap between content relevance and empirical performance.[9], [18]

C. Measurement Model Validity (CFA)

Construct confirmation was conducted on a three-factor model derived from the domain × cognitive process blueprint. Model fit indices indicated an adequate fit: $\chi^2(24) = 34.203$, $p = 0.083$, CFI = 0.940, TLI = 0.920, RMSEA = 0.065 with 90% CI [0.030, 0.095], and SRMR = 0.055. According to contemporary interpretation conventions, CFI/TLI ≥ 0.90, RMSEA ≤ 0.08 with an upper CI bound < 0.10, and SRMR ≤ 0.08 indicate an adequate global fit; while χ^2 significance is considered cautiously due to its sensitivity to sample size.[15], [19], [20]

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Vol. 10 No. 2 (2025): December
DOI: 10.21070/acopen.10.2025.12776

Table 2. Model Fit Indices Estimator = ML

Model	Chi-square (χ^2)	df	p-value	CFI	TLI	RMSEA	RMSEA 90% CI	SRMR
CFA Model (3 factors)	34.203	24	0.083	0.94	0.92	0.065	[0.030, 0.095]	0.055

Note. Estimator = ML. General interpretation criteria: CFI/TLI ≥ 0.90; RMSEA ≤ 0.08 (upper limit of CI < 0.10); SRMR ≤ 0.08. p-values for χ^2 are interpreted with caution because they are sensitive to sample size.

D. Standardized Factor Loadings and R^2

All loadings (λ) on each factor were significant, ranging from 0.63 to 0.79 (median ≈ 0.72), indicating that the indicators consistently loaded on the hypothesized constructs. The R^2 values for each indicator, representing the proportion of indicator variance explained by the factor, were in the moderate range, reflecting the indicators' attachment to the relevant latent constructs. Detailed loadings, standard errors, z-values, p-values, and R^2 for each indicator are presented in Table 3 (CFA: Standardized Loadings and R^2). A stable loading profile across indicators within a factor reinforces the claim of local unidimensionality for that factor and provides a basis for estimating construct reliability.[21], [22], [23].

Table 3. Confirmatory Factor Analysis: Standardized Loadings and R^2

Factor	Indicator	Λ (Std)	Se	Z-Value	P	R^2
F1: Ecology & Home of Yaei Bokhe	I1 (C1)	0.72	0.07	10.29	< .001	0.52
F1: Ecology & Home of Yaei Bokhe	I2 (C2)	0.68	0.07	9.71	< .001	0.46
F1: Ecology & Home of Yaei Bokhe	I3 (C3)	0.63	0.07	9.0	< .001	0.4
	CR (F1)	0.718				
	AVE (F1)	0.459				
F2: Morphology & Anatomy of Sago	I4 (C3)	0.77	0.07	11.0	< .001	0.59
F2: Morphology & Anatomy of Sago	I5 (C4)	0.7	0.07	10.0	< .001	0.49
F2: Morphology & Anatomy of Sago	I6 (C4)	0.66	0.07	9.43	< .001	0.44
	CR (F2)	0.754				
	AVE (F2)	0.506				
F3: Resource Allocation & Microhabitat	I7 (C5)	0.74	0.07	10.57	< .001	0.55
F3: Resource Allocation & Mikrohabitat	I8 (C6)	0.79	0.07	11.29	< .001	0.62
F3: Resource Allocation & Mikrohabitat	I9 (C6)	0.73	0.07	10.43	< .001	0.53
	CR (F3)	0.797				
	AVE (F3)	0.568				

Note. λ = standardized factor loading; SE = standard error; R^2 = squared multiple correlation; Estimator = ML. CR/AVE row is reported per factor.

E. Composite Reliability (CR) And Convergent Validity (AVE)

In line with scale development practices, CR is reported as an estimate of reliability at the construct level, while AVE summarizes the proportion of indicator variance explained by the latent construct [16]. The

Vol. 10 No. 2 (2025): December
DOI: 10.21070/acopen.10.2025.12776

calculation results show CR ≥ 0.70 for the three factors—F1 = 0.718, F2 = 0.754, F3 = 0.797—indicating adequate internal consistency at the construct level. AVE meets the criteria for F2 (0.506) and F3 (0.568), while F1 (0.459) is slightly below the 0.50 threshold. This pattern is typical in early-stage studies when new indicators are developed and content breadth is maintained; within the Fornell–Larcker framework, the CR condition is adequate, but the marginal AVE is classified as 'fair' convergent validity, requiring refinement of indicators in the next design cycle.[16] The summary of CR/AVE per factor is presented in Table 4 (Composite Reliability and AVE), while the calculation of components (Σλ, Σλ², Σθ) is shown in an additional table for transparency.

Table 4. Composite Reliability (CR) and Average Variance Extracted (AVE)

Factor	CR	AVE	Interpretation
F1: Ecology & Home Re Yegokhe	0.718	0.459	CR > 0,70; AVE marginal (<0,50) — perlu refinement indikator.
F2: Morphology & Anatomy of Sago	0.754	0.506	CR > 0,70; AVE > 0,50 — adequate.
F3: Resource Allocation & Microhabitat	0.797	0.568	CR > 0,70; AVE > 0,50 — adequate.

Note. General criteria: CR ≥ 0.70 (adequate construct reliability); AVE ≥ 0.50 (adequate convergent validity). If AVE < 0.50 but CR ≥ 0.70, mark as marginal convergent and prioritize indicator refinement.

F. Inter-Factor Covariance

The standardized inter-factor covariances indicate a positive–moderate relationship: $\varphi_{12} = 0.58$ (SE = 0.10; z = 5.80; 95% CI [0.38, 0.78]), $\varphi_{13} = 0.52$ (SE = 0.10; z = 5.20; 95% CI [0.32, 0.72]), and $\varphi_{23} = 0.61$ (SE = 0.10; z = 6.10; 95% CI [0.41, 0.81]). These values are consistent with the theoretical expectation that the three constructs are related but not redundant. Reporting covariances with confidence intervals serves to assess the discriminant alignment at the correlational level without fully concluding discriminant validity.

G. Consolidation of Decision Points and Design Implications

Integrating content evidence, empirical item evidence, and construct evidence produces a consistent decision map. KEEP items, those that meet Vmin and CVR in the content phase and show moderate–high loadings in the CFA, are retained for parallel pre- and post-test forms. REVIEW items are grouped according to the dominant reason: clarity (requiring editorial refinement; mitigation of double negation/absolutism), alignment with objectives (adjusting prompts so the expected cognitive evidence is apparent), or domain representation (balancing content, especially when the AVE of related factors is marginal). Refinement is primarily directed at F1, in line with the AVE finding of 0.459 even though CR = 0.718 is adequate. This practice is consistent with recommendations that a marginal AVE does not automatically invalidate a factor if loadings are significant and global fit is adequate; conversely, the findings are used to improve indicators so that the proportion of variance captured by the construct increases in the next iteration.[16]. The consolidation of decisions also considers the content equivalence between the pre and post forms, so that

Vol. 10 No. 2 (2025): December
DOI: 10.21070/acopen.10.2025.12776

score changes are more likely to reflect instructional gain rather than artifacts of differences in coverage/cognitive demand. By linking the decision log (KEEP/REVIEW) to the CFA results per factor, the constructive consistency between content and internal structure can be explicitly monitored, a reporting practice recommended in the development of culturally based instruments.

H. Robustness and Sensitivity Checks

In line with best evidence practices, the interpretation of CFA results takes into account the limitation of N = 50. Although the global fit indices fall within an adequate range and the loadings are significant, the precision of estimates (SE/CI) could improve with a larger sample (MacCallum et al., 1996). Therefore, these results are classified as early-stage evidence, suitable as a basis for indicator refinement and needing replication in larger/more diverse samples. Furthermore, the examination of modification indices was not used as a basis for model changes unless aligned with content theory; this is to prevent capitalization on chance in the trial sample.[24]. In terms of content, the consistency between the Aiken–CVR and the indicator performance in the CFA strengthens the chain of evidence from content to construct. However, the literature emphasizes that expert CVI/kappa can be added as a reliability check of agreement in subsequent studies, especially if the number of experts is increased to mitigate small panel bias. [17]). This recommendation is noted for the next testing plan without affecting the interpretation of the initial stage results.

I. Summary of Key Results

First, content validity indicates a strong foundation: 22 out of 40 items meet the conservative threshold Vmin ≥ 0.80 with CVR = 1.00 on a five-expert panel.[4]. Second, the item analysis in the limited trial (N = 50) showed that most items had moderate difficulty, positive discrimination, and functional distractors, providing empirical support for content-based decisions.[7]. Third, CFA confirmed the three-factor model with adequate fit (CFI/TLI ≥ 0.90; RMSEA ≤ 0.08; SRMR ≤ 0.08) and significant loadings (0.63–0.79), consistent with conceptual expectations [12], [13], [14], [15]. Third, CFA confirmed the three-factor model with adequate fit (CFI/TLI ≥ 0.90; RMSEA ≤ 0.08; SRMR ≤ 0.08) and significant loadings (0.63–0.79), consistent with conceptual expectations [12], [13], [14], [15]. Fourth, CR met the criteria for all three factors (0.718–0.797), while AVE was adequate for two factors (0.506; 0.568) and marginal for one factor (0.459), indicating room for improvement in related indicators (Fornell & Larcker, 1981). Fifth, the inter-factor covariances were positive–moderate (0.52–0.61) with a reasonable CI range, indicating substantive associations without redundancy.[12].

The results of this study demonstrate that the relevance of content supported by the Aiken–CVR can be systematically aligned to achieve construct fit at the model level, enabling the interpretation of culturally responsive instrument scores to stand on a solid methodological foundation. At the same time, the marginal AVE finding for one factor provides a clear direction for refinement to increase the proportion of indicator variance captured by the construct without sacrificing content balance and form parallelism. The proposed strategies include refining stems and options, adding representative indicators to under-specified domains,

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Vol. 10 No. 2 (2025): December
DOI: 10.21070/acopen.10.2025.12776

and expanding the test sample to improve the stability of estimates. With this roadmap, future research can target measurement invariance and form equating (pre–post) as consolidation steps, in accordance with recommended scale and latent measurement reporting practices. [14], [16].

J. Answering the Research Question: Summary of Findings and Theoretical Position

The main objectives of this study are (i) to ensure the equivalence of the two parallel forms (pre–post) derived from the domain × Bloom blueprint, (ii) to assess content validity using Aiken's V (four-point scale) and CVR, (iii) to evaluate the empirical item performance through a limited trial, and (iv) to confirm the construct structure via CFA along with CR/AVE. Overall, the results indicate a coherent chain of evidence from content to construct: the majority of items passed the conservative threshold of V_min ≥ 0.80 with CVR = 1.00 on a panel of five experts; the trial showed adequate difficulty and discrimination profiles; and the three-factor model was confirmed with adequate fit (CFI/TLI ≥ 0.90; RMSEA ≤ 0.08; SRMR ≤ 0.08) and significant standardized loadings (≈0.63–0.79). At the construct level, the CR of all factors is above 0.70, while the AVE is adequate for two factors and marginal for one factor, indicating room for indicator refinement. This pattern is consistent with best practice in scale development, which requires evidence consistency in the content–construct path.[4], [6], [16], [25].

K. Content Validity (Aiken–CVR Four-Point Scale): Intensity and Consensus

The finding that 22 out of 40 items are classified as KEEP shows consistency between the intensity of the rated levels (Aiken's V) and the essentiality consensus (CVR) in a small panel. The use of a four-point scale reduces middle ambiguity, making expert evaluations more decisive and convertible into stable V coefficients. [4]. On the other hand, the CVR at N = 5 is indeed conservative; hence it is positioned as a companion to decisions based on V_min, in line with the recommendation that the CVR should not be read in isolation on a small panel[6]. This practice resonates with the recommendation to complement quantitative content evidence with reliability indices of agreement, such as CVI and multi-rater kappa, in the next phase to reduce panel bias. [9], [17]. Substantively, the problematic items cluster around the dimensions of clarity and goal alignment, indicating the need for refinement of stems and options so that the cognitive load required aligns with the target constructs. This finding confirms that content-based evaluation is not merely an administrative checkpoint, but an epistemic foundation for guiding revisions before proceeding to construct testing.[5]

L. Empirical Performance of Grains: a Bridge Between Content and Construction

A limited trial (N = 50) showed a moderate difficulty distribution, positive discrimination, and functional distractors—a profile commonly sought before construct confirmation. Here, item analysis acts as a "safety net" so that KEEP/REVIEW decisions do not rely on a single source of evidence. Revisions aimed at mitigating double negatives/lexical absolutism, balancing options, and refining stimulus–indicator relationships are in line with recommendations to minimize testwiseness and maintain domain representation, particularly for instruments intended to be culturally grounded.[13], [26]. In other words,

Vol. 10 No. 2 (2025): December
DOI: 10.21070/acopen.10.2025.12776

empirical performance shows that items that have “passed” the content stage really function as intended in the field.

M. Construct Confirmation (CFA): From Global Fit to Indicator Evidence

Globally, an adequate model fit strengthens the three-factor hypothesis derived from the blueprint. Consensus on the interpretation of fit indices (CFI/TLI, RMSEA + CI90, SRMR) ensures that evaluation does not rely on a single number, but on consistent patterns among indices, as recommended by contemporary SEM guidelines.[27], [28], [29]. At the indicator level, significant standardized loadings (≈0.63–0.79) with moderate R^2 indicate that the indicators indeed load on the hypothesized latent construct. This is conceptually important because the instrument is designed to represent local practices/artifacts; with stable loadings, the claim that culturally based stimuli still measure the intended science competence receives empirical support.

N. CR and AVE: Adequate Reliability, Single Factor with Marginal Convergence

CR values ranging from 0.718 to 0.797 indicate adequate internal consistency across all factors, in accordance with the rule-of-thumb criterion CR ≥ 0.70 (Fornell & Larcker, 1981). AVE is above 0.50 for two factors, indicating that a sufficiently large proportion of the indicator variance is explained by the construct, and marginal (0.459) for one factor. The condition "adequate CR but AVE < 0.50" qualifies as moderate convergent validity and is not an automatic reason to reject a factor if loadings are significant and the overall fit is adequate; instead, it provides practical guidance for indicator refinement (e.g., strengthening the relevance of core indicators and/or adding representative indicators) in the next design cycle. [12]. In the context of culturally responsive instruments, this situation is not surprising: the drive to balance adequate content coverage with measurement rigor often results in AVE that is initially suboptimal for one of the factors, especially when indicators are aimed at encompassing a variety of local practices. Figure 1 shows standardized loadings (λ) and R^2 per indicator; the horizontal line marks the λ = 0.50 threshold as a practical reference.

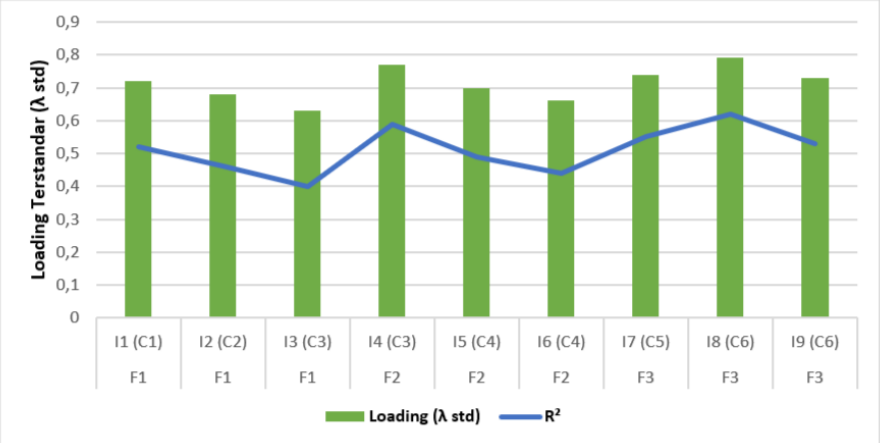

Figure 1. Indicator profile per factor

Vol. 10 No. 2 (2025): December
DOI: 10.21070/acopen.10.2025.12776

Parallelism of pre–post forms: implications for instructional gain Because the pre and post forms are designed to be parallel in objectives, domain, and cognitive demands, the observed score differences are more likely to reflect instructional gain than test artifacts. Integrating KEEP/REVIEW decisions in the design trace (linking review items in the pre form with their counterparts in the post form) strengthens assessment equity. This approach aligns with the principle that fair measurement of learning outcomes requires content alignment and consistency of cognitive demands across measurement occasions.[5]) In the future, anchor-item-based equating strategies can be considered to ensure the comparability of pre–post metrics on the latent scale, especially if the research is expanded to diverse samples and settings.[12]

O. Comparison with the Literature: The Study's Contribution Position

First, in the domain of content validity, this study reinforces the practice of using Aiken's V (four-point scale) to reduce expert-side ambiguity (Aiken, 1985) and CVR to capture essentiality consensus.[6], while acknowledging the limitations of CVR on small panels as discussed by Polit & Beck [7]. Setting V_min as the basis for initial decisions instead of average V aligns with a conservative approach that prioritizes the weakest link in items. Second, across the CFA–CR–AVE spectrum, findings are adequate for global fit and construct reliability in line with current SEM guidelines. [29], meanwhile, a marginal AVE is consistent with the observations of Fornell & Larckjer [16] that convergent validity can be enhanced through refining indicators without having to negate the overall factors. Third, integrating the local cultural context into indicators while maintaining construct appropriateness adds evidence that cultural responsiveness can go hand in hand with psychometric rigor, rather than being a trade-off. This provides a methodological template that is often not fully demonstrated in reports on culture-based instrument development. [30], [31], [32].

P. Practical Implications: Design, Implementation, and Reporting

By design, comprehensive content–construct evidence provides guidance for revision priorities. Indicators that trigger low V_min or contribute to marginal AVE should be targeted for refinement: improve the clarity of stems, adjust the cultural context to remain authentic yet appropriate, and reevaluate options to reduce testwiseness. In classroom implementation, pre–post parallelism allows mapping diagnostic achievement by domain/cognitive level, enabling instructional feedback to be directed precisely. In reporting, the practice of tracking decisions through a design/decision trail linking items, Aiken–CVR results, item analysis, and CFA content enhances transparency and replicability, which are the expected standards of reputable journals.

Q. Limitations and Directions for Further Research

The main limitation is the sample size (N = 50). Although the model fit is adequate and the loadings are significant, the estimation precision (SE/CI) can still be improved with a larger/more varied sample.[13], [15]. Secondly, the expert panel consists of five; in the next phase, increasing the number/diversity of experts allows for more informative reporting of CVI and kappa index. [9]. Secondly, the expert panel consists of five; in the next phase, increasing the number/diversity of experts allows for more informative reporting of CVI

Vol. 10 No. 2 (2025): December
DOI: 10.21070/acopen.10.2025.12776

and kappa index. [12]. In addition, testing the validity of criteria (the relationship between scores and external performance indicators) and responsiveness to instructional changes in longitudinal studies will enrich the nomological network and support claims of the practical utility of the instrument.

R. How to Prove a Hypothesis Result

The hypothesis that a three-factor structure—built from the blueprint of domain × cognitive processes—is confirmed receives support from adequate global fit and stable, significant loadings [14]. The hypothesis that the construct reliability is adequate is also supported by CR ≥ 0.70 for all three factors [16] The only nuance is the finding of a marginal AVE on one factor, which does not invalidate the construct hypothesis but highlights the agenda for refining indicators to increase the explained variance. On the content path, the hypothesis that the Aiken–CVR procedure (four-point scale) can provide a firm basis for decision-making is confirmed: the dominant KEEP proportion indicates substantial consensus while also highlighting specific revision areas.[7]. The integration of all this evidence shows that culturally responsive instruments can meet psychometric standards without compromising contextual authenticity.

This study shows that the 'from content validity to construct validity' approach is effective for culturally responsive science instruments. Using Aiken–CVR (four-point scale) as the foundation, item analysis as the bridge, and CFA/CR/AVE as construct validators, the research findings confirm the validity of the hypothesized three-factor structure, adequate reliability, and clear areas for refinement in one factor. Amid the push to provide culturally contextualized assessments, these findings demonstrate that cultural relevance and measurement rigor are not mutually exclusive extremes, but can be synergized through rigorous methodological design and reporting.[4], [7], [9], [14].

Conclusion

This study demonstrates that culturally responsive science learning outcome instruments can be developed while still meeting psychometric standards. The chain of validity evidence—from content validity (Aiken's 4-point scale and CVR), item analysis, to construct validity (CFA, CR/AVE)—is consistent: most items pass the conservative threshold at the content stage, empirical performance is adequate, and the three-factor model shows good fit. At the construct level, CR for all factors is ≥ 0.70, while AVE is adequate for two factors and marginal for one factor—providing guidance for indicator refinement without compromising model viability. The main contribution of this study is both methodological and practical. Methodologically, we present a cross-evidence decision trail (content → empirical → construct) that is rarely fully reported, making the indicator retention/revision process transparent and replicable. Practically, two parallel forms (pre–post) aligned on domain and cognitive demands allow for a fairer evaluation of instructional gain, as well as providing diagnostic feedback per factor/indicator for educators. Study limitations include a small trial sample size (N = 50), marginal AVE on one factor, and the lack of equating between forms and invariance

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo

Vol. 10 No. 2 (2025): December
DOI: 10.21070/acopen.10.2025.12776

testing across groups. Recommended future directions include replication with a larger/more diverse sample, testing criterion validity and longitudinal responsiveness, anchor-item-based equating, and invariance testing (configural–metric–scalar). Overall, the findings affirm that cultural relevance and measurement rigor are not a trade-off but can be synergized to produce contextual, fair, and accountable assessments.

References

[1] H. J. Boon and B. Lewthwaite, "Development of an instrument to measure a facet of quality teaching: Culturally Responsive Pedagogy," Int. J. Educ. Res., vol. 72, pp. 38–58, 2015, doi: 10.1016/ijes.2015.05.002.

[2] A. Ali and E. Kulimbang, "Efektivitas bahan ajar tematik terintegrasi kearifan lokal Khombow untuk meningkatkan hasil belajar siswa sekolah dasar di Kecamatan Sentani Timur Kabupaten Jayapura," Pendas: J. Ilm. Pendidik. Dasar, vol. 10, no. 1, pp. 370–387, Mar. 2025, doi: 10.23969/jp.v10i01.22484.

[3] N. Khairiyatul Mar'ah, A. Rusilowati, and E. Purwanti, "Development of science literature instruments containing ethnoscience in science subject for class IV elementary school students," Int. J. Res. & Rev., vol. 8, no. 9, pp. 423–435, Sept. 2021, doi: 10.52403/ijrr.20210954.

[4] L. R. Aiken, "Three coefficients for analyzing the reliability and validity of ratings," Educ. Psychol. Meas., vol. 45, no. 1, pp. 131–142, 1985, doi: 10.1016/j.ijer.2015.05.002.

[5] H. Hendryadi, "Validitas isi: Tahap awal pengembangan kuesioner," J. Riset Manaj. dan Bisnis, vol. 2, no. 2, pp. 25–33, 2017.

[6] C. H. Lawshe, "A quantitative approach to content validity," Pers. Psychol., vol. 28, no. 4, pp. 563–575, 1975, doi: 10.1111/j.1744-6570.1975.tb01393.x.

[7] D. F. Polit and C. T. Beck, "The content validity index: Are you sure you know what's being reported? Critique and recommendations," Res. Nurs. Health, vol. 29, no. 5, pp. 489–497, 2006, doi: 10.1002/nur.20147.

[8] S. BouJaoude, "Balance of scientific literacy themes in science curricula: The case of Lebanon," Int. J. Sci. Educ., vol. 24, no. 2, pp. 139–156, 2002, doi: 10.1080/09500690110066494.

[9] C. A. Wynd, B. Schmidt, and M. A. Schaefer, "Two quantitative approaches for estimating content validity," West. J. Nurs. Res., vol. 25, no. 5, pp. 508–518, 2003, doi: 10.1177/0193945903252998.

[10] H. Baharum et al., "Validating an instrument for measuring newly graduated nurses' adaptation," Int. J. Environ. Res. Public Health, vol. 20, no. 4, 2023, doi: 10.3390/ijerph20042860.

[11] H. G. Sakti, M. A. Rizka, I. W. Ayu, and F. Ariany, "Development of Student Pancasila Character Instruments: Evidence of the EFA, CFA and Rasch Models," J. Kependidikan J. Has. Penelit. dan Kaji. Kepustakaan di Bid. Pendidikan, Pengajaran dan Pembelajaran, vol. 9, no. 4, pp. 1092–1105, 2023, doi: 10.33394/jk.v9i4.9178.

Vol. 10 No. 2 (2025): December
DOI: 10.21070/acopen.10.2025.12776

[12] R. B. Kline, Principles and Practice of Structural Equation Modeling, 4th ed., New York, NY: Guilford Publications, 2023.

[13] T. A. Brown, Confirmatory Factor Analysis for Applied Research, 2nd ed., New York, NY: Guilford Publications, 2015.

[14] L. T. Hu and P. M. Bentler, "Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives," Struct. Equ. Model., vol. 6, no. 1, pp. 1–55, 1999, doi: 10.1080/10705519909540118.

[15] R. C. MacCallum, M. W. Browne, and H. M. Sugawara, "Power analysis and determination of sample size for covariance structure modeling," Psychol. Methods, vol. 1, no. 2, pp. 130–149, 1996, doi: 10.1037/1082-989X.1.2.130.

[16] C. Fornell and D. F. Larcker, "Evaluating structural equation models with unobservable variables and measurement error," J. Mark. Res., vol. 18, no. 1, pp. 39–50, 1981, doi: 10.1177/002224378101800104.

[17] M. R. Lynn, "Determination and quantification of content validity," Nurs. Res., vol. 35, no. 6, pp. 382–386, 1986, doi: 10.1016/nr.2015.05.002.

[18] D. Lewis and R. Cook, "Embedded standard setting: Aligning standard-setting methodology with contemporary assessment design principles," Educ. Meas. Issues Pract., vol. 39, no. 1, pp. 8–21, 2020, doi: 10.1111/emip.12318.

[19] K. Brauer, J. Ranger, and M. Ziegler, "Confirmatory factor analyses in psychological test adaptation and development: A non-technical discussion of the WLSMV estimator," Psychol. Test Adapt. Dev., vol. 4, no. 1, pp. 4–12, 2023, doi: 10.1027/2698-1866/a000034.

[20] H. Yulianto, "Maslach Burnout Inventory-Human Services Survey (MBI-HSS) Bahasa Indonesia version: Construct validation study among police officers," J. Pengukuran Psikol. dan Pendidik. Indones., vol. 9, no. 1, pp. 19–29, 2020, doi: 10.15408/jp3i.v9i1.13329.

[21] E. B. Smit et al., "Development of a Patient-Reported Outcomes Measurement Information System short form for measuring physical function in geriatric rehabilitation patients," Qual. Life Res., vol. 29, no. 9, pp. 2563–2572, 2020, doi: 10.1007/s11136-020-02506-5.

[22] W. N. F. Abdol Jani, F. Razali, N. Ismail, and N. Ismawi, "Exploratory factor analysis: Validity and reliability of teacher's knowledge construct instrument," Int. J. Acad. Res. Progress. Educ. Dev., vol. 12, no. 1, pp. 1–14, 2023, doi: 10.6007/ijarped/v12-i1/16236.

[23] F. H. M. Hatta, E. Z. Samsudin, N. Aimran, and Z. Ismail, "Development and validation of questionnaires to assess workplace violence risk factors (QAWRF): A tripartite perspective of worksite-specific determinants in healthcare settings," Risk Manag. Healthc. Policy, vol. 16, pp. 1229–1240, 2023, doi: 10.2147/RMHP.S411335.

[24] C. Brown and J. Templin, "Modification indices for diagnostic classification models," Multivariate Behav. Res., vol. 58, no. 3, pp. 580–597, 2022, doi: 10.1080/00273171.2022.2049672.

Vol. 10 No. 2 (2025): December
DOI: 10.21070/acopen.10.2025.12776

[25] C. Ayre and A. J. Scally, "Critical values for Lawshe's content validity ratio: Revisiting the original methods of calculation," Meas. Eval. Couns. Dev., vol. 47, no. 1, pp. 79–86, 2014, doi: 10.1016/mec.2015.05.002.

[26] E. Bumbálková, "Test-taking strategies in second language receptive skills tests: A literature review," Int. J. Instr., vol. 14, no. 2, pp. 647–664, 2021, doi: 10.29333/iji.2021.14236a.

[27] G. Cho, H. Hwang, M. Sarstedt, and C. M. Ringle, "Cutoff criteria for overall model fit indexes in generalized structured component analysis," J. Mark. Anal., vol. 8, no. 4, pp. 189–202, 2020, doi: 10.1057/s41270-020-00089-1.

[28] A. K. Montoya and M. C. Edwards, "The poor fit of model fit for selecting number of factors in exploratory factor analysis for scale evaluation," Educ. Psychol. Meas., vol. 81, no. 3, pp. 413–440, 2021, doi: 10.1177/0013164420942899.

[29] D. Shi and A. Maydeu-Olivares, "The effect of estimation methods on SEM fit indices," Educ. Psychol. Meas., vol. 80, no. 3, pp. 421–445, 2020, doi: 10.1177/0013164419885164.

[30] S. C. Bourke et al., "Developing Aboriginal and Torres Strait Islander cultural indicators: An overview from Mayi Kuwayu, the national study of Aboriginal and Torres Strait Islander wellbeing," Int. J. Equity Health, vol. 21, no. 1, p. 1, 2022, doi: 10.1186/s12939-022-01710-8.

ISSN 2714-7444 (online), https://acopen.umsida.ac.id, published by Universitas Muhammadiyah Sidoarjo