International Journal of Evaluation and Research in Education (IJERE)
Vol. 15, No. 6, December 2026, pp. 4830~4842
ISSN: 2252-8822, DOI: 10.11591/ijere.v15i6.39691

Method effects of negative items on mathematics self-confidence: confirmatory factor analysis of Indonesian pre-service teachers

Davi Apriandi[1], Heri Retnawati[2], Agus Maman Abadi[2], Annis Deshinta Ayuningtyas[3]
[1]Faculty of Teacher Training and Education, Universitas PGRI Madiun, Madiun, Indonesia
[2]Faculty of Mathematics and Natural Sciences, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia
[3]Faculty of Teacher Training and Education, Universitas Sarjanawiyata Tamansiswa, Yogyakarta, Indonesia

Article Info

Article history:

Received Mar 27, 2026
Revised Jul 31, 2026
Accepted Aug 6, 2026

Keywords:

Construct validity
Method effect
Negatively worded items
Pre-service mathematics teachers
Self-confidence

ABSTRACT

Negatively worded items remain common in psychological scales, despite evidence of their negative impact. This study developed a mathematics self-confidence scale for Indonesian pre-service teachers and examined how negatively worded items affect factor structure in a non-Western context. Participants included 175 pre-service mathematics teachers from six Indonesian universities. The initial 24-item model, with 16 favorable and 8 unfavorable items, showed poor fit (Chi-square/degrees of freedom (χ^2/df)=3.363; root mean square error of approximation (RMSEA)=0.117; comparative fit index (CFI)=0.720). All 8 unfavorable items had very low standardized factor loadings (λ=0.09–0.39), consistent with systematic method effects. After removing these items and respecifying the model, the final 16-item version showed good fit (χ^2/df=1.96; RMSEA=0.074; CFI=0.92), with construct reliabilities ranging from 0.70 to 0.85 and a Cronbach's alpha of 0.905. These results may relate to cognitive processing demands, Indonesian negation structures, and collectivist cultural response patterns. Due to the single-sample design, post hoc modifications, and the moderate sample size, these findings should be interpreted with caution. The study offers evidence-based guidance rather than definitive rules, supporting the use of positively worded items to measure affective constructs in similar cultural contexts, pending further replication.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Davi Apriandi
Faculty of Teacher Training and Education, Universitas PGRI Madiun
85 Setiabudi Street, Madiun 63118, Indonesia
Email: davi.mathedu@unipma.ac.id

1. INTRODUCTION

Inaccurate measurement of teachers' self-confidence has significant implications for educational research and practice. Distorted scores from psychometric instruments can result in ineffective teacher development programs being misclassified as successful, in the discontinuation of valuable interventions due to perceived null effects, and in policy decisions regarding curriculum reform or teacher certification being based on flawed evidence. In Indonesia, where more than 300,000 pre-service mathematics teachers enroll in teacher education programs annually, even minor measurement errors can lead to substantial misallocation of professional development resources. Thus, maintaining the psychometric integrity of affective instruments is essential for evidence-based educational policy.

Self-confidence is a psychological construct that plays a central role in students' academic success, particularly within mathematics education [1], [2]. Although self-confidence and self-efficacy are related,

they are conceptually distinct. Self-efficacy refers to an individual's belief in their ability to perform specific tasks or achieve particular goals [1], whereas self-confidence represents a broader dispositional belief in overall competence and capacity to act effectively across situations [3]. Stankov *et al*. [3] empirically demonstrated that self-confidence is a stronger predictor of academic achievement than self-efficacy, self-concept, or anxiety, as it reflects a more generalized sense of personal capability. In mathematics education, self-confidence encompasses students' beliefs in their ability to understand mathematical concepts, solve problems, and communicate mathematical ideas [4], [5]. Prospective mathematics teachers with high self-confidence are more likely to explore problem-solving strategies, participate actively in class discussions, and be better prepared for future pedagogical challenges [6], [7]. Consequently, valid and reliable measurement of self-confidence is essential in research on mathematics teacher education.

The development of psychological instruments has traditionally involved combining favorable (positively worded) and unfavorable (negatively worded) items to reduce acquiescence bias and response sets [8]. This method assumes that respondents who indiscriminately select "agree" can be identified through inconsistencies between responses to favorable and unfavorable items. However, increasing empirical evidence indicates that negatively worded items may introduce significant psychometric problems [9], [10].

Weijters *et al*. [9] demonstrated that negatively worded items on Likert scales are frequently associated with method effects that alter factor structure. These effects are often attributed to negatively worded items, which create artifactual factors that do not represent substantive constructs but rather reflect item-writing formats [11], [12]. Kam and Sun [13] argued that the method factor observed with reverse-keyed items may capture genuine differences in how respondents process item content, rather than mere response-style artifacts. Dueber *et al*. [14] confirmed that reverse-scored items contribute substantial method variance in affective instruments, often outweighing the intended benefits of mixed-format scales. Additionally, Salazar [15] found that negatively worded items are not consistently effective at reducing acquiescence bias and may negatively impact a scale's internal consistency.

Recent research has extended these findings across various instruments and populations. Dodeen [16] experimentally compared scales with and without negatively worded items among university students, finding that replacing negative items with positively reworded equivalents improved both reliability and factor structure. Zeng *et al*. [17] used item response theory (IRT) to analyze four versions of a learning burnout scale with different item wording valences, demonstrating that mixing positive and negative items decreased measurement reliability and introduced method factors that complicated latent trait estimation. Baumgartner *et al*. [18] developed a conceptual framework distinguishing three forms of misresponse—reversal, negation, and polar opposite—and empirically showed that processing difficulty, rather than inattention, may drive misresponse to reversed items. Collectively, these studies indicate that the detrimental impact of negatively worded items is evident across instruments, methods, and populations.

Despite a robust evidence base, most research on the method effects of negatively worded items has focused on Western, English-speaking populations [19], [20]. Studies involving non-Western populations, particularly in Southeast Asia, are limited. This gap is significant because linguistic and cultural factors may amplify the impact of negatively worded items [21]. Although method effects are well documented in Western contexts, their influence in non-Western, collectivist, or linguistically diverse populations, such as those in Indonesia, remains underexplored. A small but growing body of work has begun to extend this inquiry beyond Western settings. Antoniou and Alghamdi [22] showed that reverse-coded items confounded the measurement of mathematics confidence among Saudi Arabian eighth-graders. Within Indonesia, research by Setiawati *et al*. [23] found that reverse-worded items distorted the psychometric properties of a marital quality scale. At the same time, Mudi *et al*. [24] demonstrated through bifactor modeling that negatively worded items functioned as method artifacts rather than substantive indicators in a chemistry identity scale. Nevertheless, evidence for affective mathematics constructs measured in *Bahasa Indonesia* among pre-service mathematics teachers remains scarce, leaving the gap addressed by the study largely open.

The literature proposes three potential mechanisms to explain the failure of negatively worded items. Each mechanism may be particularly relevant in the Indonesian context and is presented as a testable hypothesis rather than a definitive conclusion. Identifying these mechanisms matters because each one implies a different remedy at the item-design stage. The present study, therefore treats them as analytic lenses for interpreting the observed loading pattern rather than as competing causal claims to be adjudicated.

The first mechanism involves increased cognitive processing demands. Respondents may need to mentally reverse the meaning of negatively worded statements before mapping them onto their self-assessment, which imposes additional cognitive load [25], [26]. This burden is especially relevant for items with double negation or subtle negation structures, which are common in *Bahasa Indonesia* translations.

The second mechanism concerns linguistic features of negation in *Bahasa Indonesia*. Unlike English, which typically marks negation with auxiliary verbs or prefixes, Indonesian uses the particle "*tidak*" before verbs or adjectives. This structure can create ambiguity when combined with reverse-scored Likert scales. For instance, the statement "*Saya tidak yakin dengan kemampuan saya*" (I am not confident in my

abilities) requires respondents to disagree with a negative statement, effectively processing a double negation, which may lead to confusion and distorted responses [27].

The third mechanism concerns cultural response tendencies in collectivist societies. Previous study found that individuals in highly collectivist cultures, such as Indonesia, may exhibit a higher tendency toward acquiescence (agreeing regardless of content) and social desirability to maintain social harmony [28]. When confronted with negatively worded items, Indonesian respondents may be more inclined to agree with the surface-level wording rather than engage in the effortful cognitive reversal required for accurate self-reporting. He and Vijver [29] argued that such response-style biases are particularly pronounced in cross-cultural assessment contexts in which instruments are developed in one cultural setting and applied in another.

It is important to acknowledge that these three mechanisms are not the only plausible explanations for the failure of negatively worded items. Alternative explanations, such as item ambiguity in translation, content-specific difficulty unrelated to wording direction, or individual differences in careless responding [19], may also contribute to the observed patterns and cannot be definitively ruled out without experimental manipulation. Acknowledging these alternatives is important for bounding the interpretive scope of the findings. The analysis is therefore framed in associative rather than strictly causal terms.

The use of negatively worded items has direct implications for the validity of educational research. If method effects from such items distort factor structures, research conclusions based on these scores may be compromised. In mathematics teacher education, where affective constructs such as self-confidence are increasingly used to evaluate teaching interventions and curriculum reforms, psychometrically flawed instruments may yield inaccurate assessments of teacher confidence, misguided policy decisions, or ineffective interventions. Practically, misestimating teachers' self-confidence could affect training program design, resource allocation, and evaluation policies, ultimately failing to address the actual needs of pre-service teachers. Improved instrument design, informed by this research, can enable more accurate measurement of affective constructs in Indonesian and similar cultural contexts.

Most self-confidence instruments in the literature were developed within Western educational contexts [3]. These instruments may not demonstrate equivalent construct validity when applied to Indonesian prospective mathematics teachers, who possess distinct linguistic, cultural, and academic characteristics. The need for empirically validated instruments tailored to this context is increasingly urgent, given the critical importance of measurement quality in educational research. Based on the considerations, this study addresses two research questions:

- Does the self-confidence scale for Indonesian pre-service mathematics teachers demonstrate adequate construct validity and reliability as measured through confirmatory factor analysis (CFA)? (RQ1)
- Are negatively worded (unfavorable) items associated with systematic method effects that may affect the factor structure of the self-confidence scale in the Indonesian context? (RQ2)

This study empirically examines the impact of negative item wording on scale validity in Indonesia and provides practical guidance for designing culturally sensitive instruments. The contributions are both practical, through the development of a validated instrument for use in Indonesian mathematics teacher education, and methodological, by offering empirical evidence on method effects and exploring plausible underlying mechanisms in a non-Western context. The novelty of this study lies in being among the first to apply CFA to examine the method effects of negatively worded items on a self-confidence scale developed natively in *Bahasa Indonesia* for pre-service mathematics teachers, thereby extending method-effect evidence from Western, English-language settings to a non-Western, linguistically distinct context.

2. METHOD

2.1. Research design

A cross-sectional psychometric design was employed to develop and empirically validate a self-confidence scale. This instrument development and validation study utilized first-order CFA with four correlated latent dimensions as the primary analytical approach to assess construct validity and reliability [30], [31]. This design is appropriate for the study's aims because it allows the scale's internal structure and the behavior of individual items to be examined simultaneously within a single sample of respondents. Cross-sectional psychometric designs are widely used in instrument validation when the objective is to establish construct validity at a defined point in time rather than to track change longitudinally.

2.2. Participants and sampling

The study sample comprised 175 pre-service mathematics teachers (49 males, 126 females) enrolled in mathematics education programs at six universities or Teacher Training and Education Institutes (LPTK) across Indonesia. The participating institutions were located throughout the Java region, ensuring geographic diversity within Indonesia's most populous educational corridor. Convenience sampling was used: students

enrolled in the mathematics education program during the even semester of the 2024/2025 academic year were invited to participate, and only those who provided informed consent were included. Inclusion criteria required: i) active enrollment in the mathematics education study program and ii) completion of at least one semester of coursework. No additional exclusion criteria were applied.

A sample size of 175 respondents was deemed adequate for CFA, as it exceeded the minimum of 150 respondents recommended for measurement model testing [30], met the minimum requirement of five respondents per parameter [32], and aligned with the recommended ratio of 5–10 respondents per item [32]. Simulation studies have confirmed that parameter estimation in psychometric models is generally stable at sample sizes of approximately 150–200 when factor loadings are moderate to high [31]. Although this sample size surpasses traditional minimums, it remains within the lower-to-moderate range for CFA. Monte Carlo simulations indicate that required sample sizes for structural equation modeling (SEM) can vary widely, from 30 to 460 cases, depending on model complexity and parameter magnitudes [32].

Ethical procedures were followed throughout data collection. All participants were informed about the purpose of the study, the voluntary nature of participation, the anonymity of responses, and their right to withdraw at any time without consequence. Written informed consent was obtained from all participants before data collection commenced. The study was conducted in accordance with the ethical standards of the participating universities' institutional research committees, and all procedures complied with national regulations on research involving human subjects.

2.3. Instrument development

The self-confidence scale was developed specifically for this study based on a synthesis of theoretical frameworks from Bandura [1] and Stankov *et al*. [3]. The theoretical framework distinguishes self-confidence from self-efficacy by operationalizing it as a broader dispositional belief encompassing four dimensions: i) belief in self-efficacy; ii) independence and courage to act; iii) emotional calmness and optimism; and iv) self-acceptance and positive social relationships. These dimensions were derived from a comprehensive literature review and adapted to the specific context of Indonesian pre-service mathematics teachers.

All items were newly developed in *Bahasa Indonesia* rather than translated from existing scales. This native-development approach was chosen to avoid translation artifacts and ensure conceptual fidelity to Indonesian linguistic and cultural contexts. As a result, formal translation and back-translation procedures were not applicable. The absence of cognitive interviews during item development is recognized as a limitation, as such interviews could have provided valuable insight into how respondents process the items, particularly the unfavorable ones.

In the initial stage, items were newly developed (not adapted from existing scales) using a deductive approach. Each dimension was represented by six items, consisting of four favorable (positively worded) and two unfavorable (negatively worded) items, totaling 24 statements. Favorable items were written as direct affirmations of the construct ("I feel confident when facing new material in mathematics"), while unfavorable items expressed the opposite direction ("I feel anxious when asked to solve problems in front of the class"). All items were written in *Bahasa Indonesia*, as this is the primary language of instruction in the target population. The distribution of items across dimensions is presented in Table 1.

Respondents rated on a 5-point Likert scale ranging from "strongly disagree" (score 1) to "strongly agree" (score 5), with reverse scoring applied to unfavorable items before analysis. This scoring direction ensured that higher composite scores consistently represented higher self-confidence across all items. The five-point format was selected to balance response discrimination against respondent cognitive load.

Table 1. Self-confidence questionnaire grid

Dimension	Favorable	Unfavorable	Σ favorable	Σ Unfavorable	Total
A. Belief in self-efficacy	1, 2, 4, 6	3, 5	4	2	6
B. Independence and courage to act	7, 8, 10, 12	9, 11	4	2	6
C. Emotional calmness and optimism	13, 14, 16, 18	15, 17	4	2	6
D. Self-acceptance and positive social relationships	19, 20, 22, 24	21, 23	4	2	6
Total			16	8	24

2.4. Content validation

Before empirical testing, all 24 items underwent content validation by four expert validators with the following qualifications: two professors of mathematics education with expertise in psychometrics and scale development, one associate professor of educational evaluation, and one professor of educational psychology. Each validator independently rated every item on a 4-point scale (1=not relevant, 2=somewhat relevant, 3=quite relevant, 4=highly relevant) based on the clarity, relevance, and representativeness of each

item relative to its target dimension. The ratings were then analyzed using Aiken's V index [33], with a minimum threshold of V≥0.80 indicating acceptable content validity [34], [35]. All 24 items met this criterion, confirming adequate content validity for the initial item pool. Qualitative feedback from the validators was also incorporated, resulting in minor revisions to item wording for clarity before data collection.

2.5. Data analysis

To empirically test construct validity, data analysis was conducted using LISREL 8.80, employing first-order CFA with four correlated latent dimensions and the maximum likelihood (ML) estimation method. In this model, the 16 observed indicators (items) loaded onto four correlated first-order latent factors representing the four self-confidence dimensions: i) belief in self-efficacy; ii) independence and courage to act; iii) emotional calmness and optimism; and iv) self-acceptance and positive social relationships. The covariance paths among the four latent factors reflect the theoretical assumption that these dimensions are related but distinct facets of the self-confidence construct [30], [31].

Bifactor models and multitrait-multimethod (MTMM) approaches represent alternative analytical strategies that can explicitly model method effects associated with the direction of item wording [13], [17]. For example, a bifactor model would specify a general self-confidence factor alongside a method factor capturing variance unique to negatively worded items. Although such approaches have theoretical appeal for partitioning substantive variance from method variance, item elimination was selected as the primary strategy for three reasons. First, all eight negatively worded items exhibited factor loadings below the minimum threshold of 0.40, indicating that they failed as indicators of their substantive dimensions and as candidates for any model that would retain them. Second, retaining items with such low loadings within a method factor risks over-parameterizing the model relative to the sample size of 175. Third, eliminating problematic items yields a parsimonious scale that is more readily interpretable and usable in subsequent research. It is recommended that future studies with larger samples explicitly test bifactor or MTMM models to validate the interpretation of the method effect further and to compare model performance across analytical strategies.

The model evaluation process was carried out in two sequential stages. The first stage focused on testing the initial model (24 items) to identify and eliminate structurally problematic indicators. The second stage retested the revised model (16 purely favorable items) to ensure the stability of the final model. The item-level validity decision was based on the standardized factor loading criterion of $\lambda \geq 0.40$ [36].

Overall, the quality of the measurement model fit was comprehensively evaluated using several goodness-of-fit indices: Chi-square/degrees of freedom (χ^2/df) ratio (<3.00), root mean square error of approximation (RMSEA) (<0.08), comparative fit index (CFI) (≥0.90), and standardized root mean square residual (SRMR) (<0.08) [30], [31]. Additional metrics, such as normed fit index (NFI) and goodness-of-fit index (GFI), were also examined, with values in the range of 0.85–0.89 classified as a moderate fit and the remaining values methodologically acceptable. These thresholds were selected because they are widely adopted benchmarks in psychometric validation and permit comparison with prior CFA studies. Reporting multiple indices, rather than relying on a single statistic, reduces the risk of drawing conclusions from any one index in isolation.

In the final stage, internal consistency was evaluated using three complementary indicators: composite reliability (CR), average variance extracted (AVE), and Cronbach's alpha. Multiple indicators were used because each captures a different aspect of reliability: CR assesses the overall reliability of a set of indicators representing a latent construct; AVE evaluates convergent validity by measuring the proportion of variance captured by the construct relative to measurement error; and Cronbach's alpha provides a widely recognized benchmark for internal consistency [37]. Constructs were deemed reliable if CR≥0.70. Given that AVE is a more conservative measure, meeting the strong CR criterion was considered sufficient evidence of convergent validity even when AVE fell below 0.50 [30].

During the respecification of the 16-item model, correlated error terms were added based on both statistical modification indices from LISREL and conceptual justification. To minimize the risk of overfitting from purely data-driven modifications, correlated errors were constrained to pairs of items that satisfied two criteria: i) statistical modification indices indicated substantial improvement in model fit and ii) the items shared overlapping content or similar phrasing within the same dimension, justifying shared variance on theoretical grounds. For example, within dimension C (emotional calmness and optimism), the correlated error between "I remain calm when faced with difficult math problems" and "I do not give up easily, even when I fail an exam" was justified because both items address emotional regulation in the face of academic adversity. This approach follows Byrne [38] recommendation that model modifications should be theoretically defensible rather than purely data-driven. However, the absence of an independent validation sample limits the ability to evaluate the generalizability of these modifications.

3. RESULTS AND DISCUSSION

3.1. Initial confirmatory factor analysis model (24 items)

In the first stage, CFA was conducted on the initial measurement model consisting of 24 items (16 favorable and 8 unfavorable). The unfavorable item scores were reversed before analysis. The analysis results showed that the initial model had poor goodness of fit: χ^2/df=3.363; RMSEA=0.117; CFI=0.720; NFI=0.63; GFI=0.71; SRMR=0.10. Almost all indices fell below the established criteria. The eight unfavorable items were reverse-scored before this analysis to ensure that all indicators shared a common scoring direction. This step ensured that any low loadings could not be attributed merely to uncorrected scoring.

Examination of the standardized factor loadings pointed to the most likely source of the model's poor fit. As presented in Table 2, a clear and consistent pattern emerged: all eight unfavorable items had very low loadings (λ=0.09–0.39), while all 16 favorable items met the minimum criterion (λ=0.42–0.83). This item-level inspection was carried out before any indicator was removed, so that the elimination decision rested on transparent empirical evidence.

A key finding is the consistent dichotomy observed: every unfavorable item failed to meet the minimum loading threshold, whereas every favorable item met this criterion. This systematic pattern indicates that the probability of all eight unfavorable items failing and all sixteen favorable items passing by chance is negligible. Such results provide strong empirical support for the interpretation that the direction of item wording, rather than item content, may be a primary contributor to the observed misfit. All eight unfavorable items exhibited low loadings, a pattern consistent with prior research suggesting that negative wording may introduce method bias [9], [14]. Based on this evidence, these items were eliminated from subsequent analyses.

This decision aligns with the recommendation of Sonderen *et al*. [39]. It is further supported by the systematic nature of the pattern, which occurred across all unfavorable items and all four dimensions, rather than reflecting random variation. Treating the pattern as systematic, rather than as a series of isolated item failures, is consistent with how method effects are conceptualized in the measurement literature. This framing also guided the subsequent interpretation of the revised model.

Table 2. Individual standardized factor loadings for all 24 items

Item	Wording direction	Dimension	Λ	Decision
1	Favorable	A	0.82	Valid
2	Favorable	A	0.80	Valid
3	Unfavorable	A	0.30	Invalid
4	Favorable	A	0.71	Valid
5	Unfavorable	A	0.30	Invalid
6	Favorable	A	0.74	Valid
7	Favorable	B	0.60	Valid
8	Favorable	B	0.65	Valid
9	Unfavorable	B	0.31	Invalid
10	Favorable	B	0.67	Valid
11	Unfavorable	B	0.36	Invalid
12	Favorable	B	0.65	Valid
13	Favorable	C	0.55	Valid
14	Favorable	C	0.42	Valid
15	Unfavorable	C	0.09	Invalid
16	Favorable	C	0.66	Valid
17	Unfavorable	C	0.40	Invalid
18	Favorable	C	0.76	Valid
19	Favorable	D	0.72	Valid
20	Favorable	D	0.83	Valid
21	Unfavorable	D	0.19	Invalid
22	Favorable	D	0.57	Valid
23	Unfavorable	D	0.33	Invalid
24	Favorable	D	0.83	Valid

3.2. Revised confirmatory factor analysis model (16 items)

Following the removal of eight unfavorable items, CFA was performed on the remaining 16 favorable items. The initial model, before respecification, demonstrated a suboptimal fit. Guided by modification indices from LISREL and the theoretical rationale, the model was respecified by adding error-variance correlations among items. Figure 1 displays the CFA modification model (standardized solution), while Table 3 summarizes the final model's goodness-of-fit results after respecification.

The improvement across all indices is substantial. RMSEA decreased from 0.117 to 0.074, CFI increased from 0.720 to 0.92, and SRMR decreased from 0.10 to 0.061. The NFI (0.86) and GFI (0.88) indicated a moderate fit, which remains methodologically acceptable given the sample size and model

complexity [31]. These improvements indicate that the systematic removal of negatively worded items addressed a primary source of model misfit, supporting the interpretation that the 16 retained items function as valid indicators of the self-confidence construct. However, because fit improvements were derived in part from post hoc model modification and cross-validation was not conducted, the stability of the final model requires confirmation in independent samples. Table 4 presents the standardized factor loadings for the final 16-item model. All items exceeded the minimum threshold of λ≥0.40, with loadings ranging from 0.41 to 0.84, confirming adequate construct validity for each retained item.

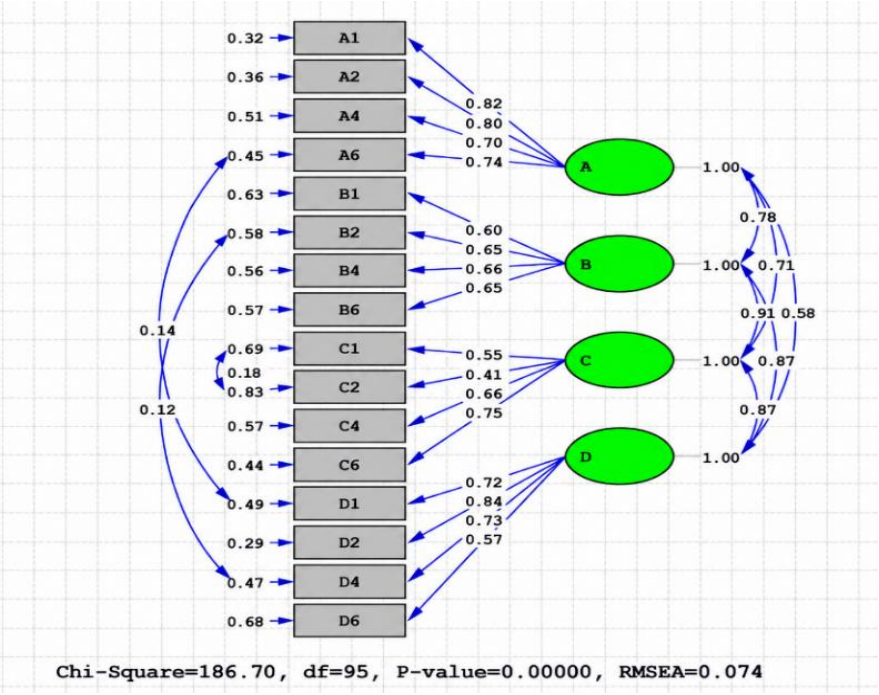


Figure 1. CFA modification model (standardized solution)

Table 3. Comparison of goodness of fit between models

Index	Criteria	Initial model (24 items)	Decision	Final model (16 items)	Decision
χ^2/df	<3	3.363	Poor fit	1.96	Good fit
RMSEA	<0.08	0.117	Poor fit	0.074	Good fit
CFI	≥0.90	0.720	Poor fit	0.92	Good fit
NFI	≥0.90	0.63	Poor fit	0.86	Moderate fit
GFI	≥0.90	0.71	Poor fit	0.88	Moderate fit
SRMR	<0.08	0.10	Poor fit	0.061	Good fit

Table 4. Standardized factor loading final model (16 items)

Dimension	Item	λ	Decision
Belief in one's own abilities	I feel confident when facing new material in mathematics	0.82	Valid
	I believe I can do math problems correctly	0.80	Valid
	I am confident that I can complete tasks without fear of failure	0.70	Valid
	I believe I have the potential to face challenges in mathematics	0.74	Valid
Independence and the courage to act	I dare to ask the lecturer when I do not understand the material	0.60	Valid
	I dare to express my opinion during group discussions	0.65	Valid
	I can complete assignments independently without cheating	0.66	Valid
	I believe in my own ability to find solutions to the mathematical problems I encounter	0.65	Valid
Emotional calmness and optimism	I remain calm when faced with difficult math problems	0.55	Valid
	I can control my emotions when exam results are not as expected	0.41	Valid
	I am optimistic that I can get good grades in this course	0.66	Valid
	I do not give up easily, even when I fail an exam	0.75	Valid
Self-acceptance and positive social relationships	I accept my abilities in learning mathematics as they are	0.72	Valid
	I am aware of my weaknesses and try to improve them	0.84	Valid
	I can work well in a study group	0.73	Valid
	I feel accepted and appreciated in the classroom environment	0.57	Valid

3.3. Composite reliability

Reliability and convergent validity of the final scale were assessed using CR, AVE, and Cronbach's alpha, as shown in Table 5. Using three complementary indices provides a more complete picture of reliability than any single coefficient. Each index addresses a distinct facet of measurement quality, allowing convergent validity and internal consistency to be evaluated together.

Table 5 indicates that all four dimensions achieved CR values above the minimum threshold of 0.70 [30], ranging from 0.70 to 0.85. Dimensions A (CR=0.85) and D (CR=0.81) exhibited strong reliability, while dimensions B (CR=0.74) and C (CR=0.70) satisfied the minimum criterion. Although the AVE values for dimensions B (AVE=0.41) and C (AVE=0.37) did not reach the conservative threshold of 0.50, this outcome is methodologically acceptable. Malhotra [40] and Nunan *et al*. [41] note that AVE is a more stringent indicator of convergent validity than CR; even if more than 50% of the variance is due to measurement error, the construct's convergent validity remains adequate when CR meets the minimum threshold.

Given these considerations, the indicators for dimensions B and C were retained. Removing items solely to meet the AVE threshold could diminish the instrument's substantive content and compromise the content validity established by the theoretical framework. Retaining these items is a methodologically justified decision to maintain the comprehensiveness of the self-confidence construct domain. This limitation is acknowledged and recommended as a focus for future instrument refinement. The overall Cronbach's alpha of 0.905 demonstrates excellent internal consistency across all 16 items [42].

Table 5. CR and AVE

Dimension	Range λ	CR	AVE	α	Decision
Belief in self-efficacy	0.70–0.82	0.85	0.59	0.843	Reliable
Independence and courage to act	0.60–0.66	0.74	0.41	0.739	Reliable
Emotional calmness and optimism	0.41–0.75	0.70	0.37	0.709	Reliable
Self-acceptance and positive social relationships	0.57–0.84	0.81	0.52	0.793	Reliable
Overall				0.905	Highly reliable

3.4. Discussion

The main finding is that all eight negatively worded items failed to meet the minimum factor-loading criterion across all four dimensions, providing empirical support for systematic method effects. This result has several implications for the literature, though we interpret these patterns as associated with item wording rather than as definitive causal claims. Framing the result in associative terms keeps the interpretation aligned with the correlational design. The discussion that follows therefore emphasizes plausible mechanisms rather than confirmed causes.

3.4.1. Consistency and systematicity of the method effect

The consistency of the observed pattern across all dimensions is noteworthy. All unfavorable items across the four dimensions exhibited the same low-loading pattern. This uniformity is more plausibly attributed to a systematic source of variance rather than to independent failures of individual items [11]. The observed pattern aligns with the integrative model proposed by Weijters *et al*. [9], empirical evidence from factor-structure testing by Zhang *et al*. [10], and recent experimental findings by Dodeen [16], who demonstrated that rewriting negatively worded items in a positive direction led to improvements in both reliability and factor structure. It should be noted, however, that "systematic" in this context refers to the consistency of the pattern across items, rather than to a confirmed causal mechanism; establishing causality would require experimental manipulation.

3.4.2. Plausible mechanisms underlying the observed pattern

A central question arising from these results concerns the consistent pattern of low loadings for negatively worded items. Although this study did not utilize an experimental design to test causal mechanisms directly, the observed findings, together with existing theoretical and empirical evidence, suggest three mutually reinforcing mechanisms. These mechanisms are proposed as hypotheses for future investigation.

Cognitive processing demands. Negatively worded items may require respondents to engage in additional cognitive operations, such as interpreting the negative meaning of a statement and then aligning their agreement with the reverse-scored scale. This dual processing can increase cognitive load, especially for items involving double negation (for example, disagreeing with a negative statement) [25], [26]. In the present study, several unfavorable items included subtle negation structures (for example, "*Saya merasa cemas ketika diminta menyelesaikan soal di depan kelas*"—I feel anxious when asked to solve problems in front of the class), which required respondents to disagree with the statement to indicate high self-confidence.

The very low factor loadings of these items (as low as λ=0.09 for item 15) support the hypothesis that many respondents did not perform this cognitive reversal accurately. However, alternative explanations such as item-specific ambiguity cannot be excluded. Linguistic features of Indonesian negation. The structure of negation in *Bahasa Indonesia* differs from English in ways that may increase processing difficulty. Indonesian negation is typically formed by placing the particle "*tidak*" before a verb or adjective, resulting in sentences that may appear grammatically simple but become semantically confusing when used in a Likert scale format [27]. Additionally, constructing negatively worded items in *Bahasa Indonesia* often requires reformulating the entire sentence, which can unintentionally alter semantic nuance or introduce unfamiliar phrasing. This potential linguistic mismatch between the intended construct and the respondent's interpretation may explain the consistently low factor loadings observed across all dimensions.

Cultural response tendencies. Indonesia is a collectivist society in which social harmony and agreeableness are highly valued [28]. In such contexts, respondents may exhibit increased acquiescence bias, a tendency to agree with statements regardless of their content [29], [43]. When presented with a negatively worded item, an acquiescent respondent may endorse the statement at face value, resulting in a response that contradicts their actual level of self-confidence, even after reverse scoring. This cultural mechanism may interact with the cognitive and linguistic mechanisms, potentially compounding the effect and reducing the reliability of negatively worded items in the Indonesian context.

Importantly, these three mechanisms are not mutually exclusive, nor do they exhaust the space of possible explanations. Alternative explanations—such as item-specific ambiguity, content overlap between negatively and positively worded items, or individual differences in careless responding [19], [20]—may also contribute to the observed pattern. Distinguishing among these explanations would require experimental designs that manipulate item wording while holding content constant, as well as cognitive interviewing methods that examine response processes directly [44].

3.4.3. Expanding the evidence base to a non-Western context

These findings extend the literature on method effects to non-Western contexts, which remain underrepresented in psychometric research. Most prior studies [9], [12], [14], [19] have focused on English-speaking populations in Western countries. The present study demonstrates that the negative impact of unfavorable items is evident and may be more pronounced in Indonesian higher education. This result aligns with recent CFA studies of mathematical instruments for Indonesian pre-service teachers, which emphasize the need for rigorous construct validation in this population [45]. These findings indicate that instrument developers should not assume that a scale developed with mixed-format items in a Western context will function equivalently in a Southeast Asian context. Cross-cultural researchers should exercise particular caution when adapting instruments containing negatively worded items for use in Indonesian or similar collectivist populations [29].

These results are consistent with previous findings from a different instrument. Apriandi *et al*. [46] validated a 17-item learning motivation questionnaire administered to mathematics education students in Indonesia. All items were positively worded and satisfied construct validity criteria, with factor loadings greater than 0.40 and a CR of 0.902. These convergent findings reinforce the argument that the observed issue is more likely attributable to the negatively worded item format itself, rather than to characteristics of the population or the analytical method.

3.4.4. Practical recommendations for instrument developers

Drawing on the findings of this study and broader international evidence [16], [20], [39], several recommendations are proposed for researchers and instrument developers working in Indonesian or comparable Southeast Asian educational contexts. These recommendations are presented as evidence-based guidance rather than prescriptive rules, acknowledging that the optimal approach may vary depending on the construct measured, the target population, and the intended application of the instrument. The recommendations should therefore be read as a starting point for context-sensitive instrument design rather than as fixed requirements.

First, when measuring affective constructs such as self-confidence among Indonesian pre-service teachers, instrument developers should prioritize positively worded items. Recent evidence challenges the assumption that negatively worded items are required to control for acquiescence bias [16]; in many cases, such items may introduce method artifacts that undermine construct validity. Prioritizing positively worded items can therefore reduce method variance while preserving the scale's substantive meaning.

Second, alternative strategies for detecting careless or acquiescent responding should be considered as substitutes for reverse-scored items. These strategies include attention checks or bogus items unrelated to the construct being assessed to identify inattentive respondents [47], analysis of response time data, and

statistical methods such as person-fit indices [30]. These approaches can serve the quality-control function traditionally assigned to reverse-scored items without introducing their associated method effects.

Third, when adapting or translating instruments that originally include negatively worded items, researchers should consider rewriting all items in a positive direction rather than translating negative formulations directly. Prior studies indicate that this approach can improve both reliability and validity [16], although context-specific validation remains necessary. Rewriting items in a positive direction is especially relevant when an instrument developed in one language is deployed in a structurally different language such as *Bahasa Indonesia*.

Fourth, if negatively worded items are retained for theoretical or practical reasons, researchers should explicitly model the method effect, for example by using a bifactor model or correlated uniqueness approach, rather than treating these items as equivalent to positively worded items [13], [17]. Explicitly modeling the method effect allows researchers to retain theoretically motivated items without conflating method variance with substantive variance. This approach is particularly valuable when comparison with prior mixed-format instruments is required.

In addition to instrument design considerations, providing training for pre-service teachers and faculty in self-assessment and survey response strategies can help mitigate potential method bias, particularly when combined with careful item design. Such training can raise respondents' awareness of how to interpret reverse-keyed statements accurately. Combined with improved item design, it offers a complementary route to reducing method bias.

3.4.5. Limitations

Several limitations of this study warrant consideration. First, although the sample was drawn from six universities, the total of 175 pre-service mathematics teachers constitutes a relatively small-to-moderate sample size for CFA. Additionally, all participants were enrolled in mathematics education programs in Java. Replication of these findings with larger samples across diverse academic disciplines, geographic regions of Indonesia, and educational levels is necessary to enhance generalizability.

Second, this study utilized a single-sample CFA without an independent validation sample. The final model fit was achieved in part through post-hoc model modification by correlating error terms. Although these modifications were theoretically justified, they introduce a risk of overfitting to the specific sample. The stability of the final model parameters across independent samples cannot be assessed with the current data. Future research should evaluate the final 16-item model in new samples using cross-validation procedures or pre-registered confirmatory designs.

Third, although this study identified and discussed three plausible mechanisms for the observed pattern of low loadings on negatively worded items (cognitive processing, linguistic structure, and cultural response tendencies), the cross-sectional correlational design precludes causal attribution. The specific mechanism or combination of mechanisms responsible cannot be definitively determined without experimental manipulation. Future research should employ experimental designs, such as randomly assigning participants to versions of the scale with and without negatively worded items, as demonstrated by Dodeen [16], to assess the relative contribution of each mechanism directly.

Fourth, the study did not incorporate a think-aloud or cognitive interview component to examine how respondents processed the negatively worded items. Qualitative data from cognitive interviews could provide valuable insights into whether respondents misunderstood, misinterpreted, or disregarded the negation in unfavorable items. Mixed-method approaches that combine CFA with qualitative cognitive interviewing are recommended for future research [44].

Fifth, alternative analytical approaches such as bifactor CFA, MTMM modeling, and IRT could offer complementary perspectives on the method effect. Although item elimination was selected for the reasons, future studies with larger samples could directly compare these approaches to more effectively distinguish substantive variance from method variance. Comparing these approaches within a single dataset would clarify how much variance is genuinely attributable to method effects. Such a comparison was beyond the scope of the present sample but represents a clear direction for future work.

4. CONCLUSION

This study contributes to the literature on instrument development in non-Western educational contexts in two primary ways. First, it introduces a self-confidence scale for Indonesian pre-service mathematics teachers, consisting of 16 positively worded items across four dimensions: belief in self-efficacy, independence and courage to act, emotional calmness and optimism, and self-acceptance and positive social relationships. The scale demonstrates strong model fit and high internal consistency (α=0.905). This instrument is suitable for research and program evaluation in Indonesian mathematics teacher education, subject to further validation in additional samples.

Methodologically, this study provides empirical evidence that negatively worded items on Likert scales may introduce systematic method effects among pre-service mathematics teachers in Indonesia. This finding extends the predominantly Western literature to a Southeast Asian educational context. An explanatory framework is proposed that involves three mutually reinforcing hypothetical mechanisms: increased cognitive processing demands associated with negation reversal, specific linguistic features of negation in *Bahasa Indonesia*, and cultural response tendencies characteristic of collectivist societies. These mechanisms are presented as testable hypotheses rather than established causal explanations, suggesting that the detrimental impact of negatively worded items may be especially pronounced in non-Western contexts. This has important implications for cross-cultural instrument development and adaptation. These findings are context-specific and should be regarded as evidence-based guidance for instrument developers working with Indonesian pre-service teachers rather than as a universal rule; they require replication across different populations, instruments, and analytical approaches before being considered definitive.

ACKNOWLEDGMENTS

The authors sincerely thank the expert lecturers for their feedback and validation of the self-confidence instrument. They also appreciate all pre-service mathematics teachers from the participating Teacher Training and Education Institutes (LPTK) in Indonesia for their time and participation as respondents. The authors also acknowledge the institutional support that made this validation study possible.

FUNDING INFORMATION

The authors state that no funding is involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Davi Apriandi	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	
Heri Retnawati	✓	✓		✓						✓		✓		
Agus Maman Abadi	✓	✓		✓						✓		✓		
Annis Deshinta Ayuningtyas		✓	✓	✓	✓			✓		✓	✓		✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, [DA], upon reasonable request.

REFERENCES

[1] A. Bandura, *Self-efficacy: the exercise of control*, 1st ed. New York, NY: W. H. Freeman and Company, 1997.
[2] D. H. Schunk and M. K. DiBenedetto, "Self-efficacy and human motivation," in *Advances in Motivation Science*, A. J. Elliot, Ed., Amsterdam: Elsevier, 2021, pp. 153–179, doi: 10.1016/bs.adms.2020.10.001.
[3] L. Stankov, J. Lee, W. Luo, and D. J. Hogan, "Confidence: a better predictor of academic achievement than self-efficacy, self-concept and anxiety?" *Learning and Individual Differences*, vol. 22, no. 6, pp. 747–758, Dec. 2012, doi: 10.1016/j.lindif.2012.05.013.
[4] P. Grootenboer and M. Marshman, *Mathematics, affect and learning: middle school students' beliefs and attitudes about mathematics education*, 1st ed. Singapore: Springer Singapore, 2016, doi: 10.1007/978-981-287-679-9.

[5] M. S. Hannula, "Exploring new dimensions of mathematics-related affect: embodied and social theories," *Research in Mathematics Education*, vol. 14, no. 2, pp. 137–161, Jul. 2012, doi: 10.1080/14794802.2012.694281.
[6] J. S. Mendías, I. S. Alex, and A. M. Espigares, "Anxiety and self-confidence toward mathematics in preservice primary education teachers," *Electronic Journal of Research in Educational Psychology*, vol. 18, no. 51, 2020, doi: 10.25115/ejrep.v18i51.2981.
[7] W. W. F. Lau, "Pre-service mathematics teachers' professional learning in a pedagogy course: examining changes in beliefs and confidence in teaching algebra," *Mathematics Education Research Journal*, vol. 33, no. 2, pp. 223–239, Jun. 2021, doi: 10.1007/s13394-019-00285-y.
[8] D. L. Paulhus, "Measurement and control of response bias," in *Measures of Personality and Social Psychological Attitudes*, J. P. Robinson, P. R. Shaver, and L. S. Wrightsman, Eds., San Diego, CA: Academic Press, 1991, pp. 17–59, doi: 10.1016/B978-0-12-590241-0.50006-X.
[9] B. Weijters, H. Baumgartner, and N. Schillewaert, "Reversed item bias: an integrative model," *Psychological Methods*, vol. 18, no. 3, pp. 320–334, 2013, doi: 10.1037/a0032121.
[10] X. Zhang, R. Noor, and V. Savalei, "Examining the effect of reverse worded items on the factor structure of the need for cognition scale," *PLoS ONE*, vol. 11, no. 6, p. e0157795, Jun. 2016, doi: 10.1371/journal.pone.0157795.
[11] P. M. Podsakoff, S. B. MacKenzie, and N. P. Podsakoff, "Sources of method bias in social science research and recommendations on how to control it," *Annual Review of Psychology*, vol. 63, no. 1, pp. 539–569, Jan. 2012, doi: 10.1146/annurev-psych-120710-100452.
[12] M. J. Roszkowski and M. Soven, "Shifting gears: consequences of including two negatively worded items in the middle of a positively worded questionnaire," *Assessment and Evaluation in Higher Education*, vol. 35, no. 1, pp. 113–130, Jan. 2010, doi: 10.1080/02602930802618344.
[13] C. C. S. Kam and S. Sun, "Method factor due to the use of reverse-keyed items: is it simply a response style artifact?" *Current Psychology*, vol. 41, no. 3, pp. 1204–1212, Mar. 2022, doi: 10.1007/s12144-020-00645-z.
[14] D. M. Dueber *et al.*, "To reverse item orientation or not to reverse item orientation, that is the question," *Assessment*, vol. 29, no. 7, pp. 1422–1440, Oct. 2022, doi: 10.1177/10731911211017635.
[15] M. S. Salazar, "The dilemma of combining positive and negative items in scales," *Psicothema*, vol. 27, no. 2, pp. 192–199, 2015, doi: 10.7334/psicothema2014.266.
[16] H. Dodeen, "The effects of changing negatively worded items to positively worded items on the reliability and the factor structure of psychological scales," *Journal of Psychoeducational Assessment*, vol. 41, no. 3, 2023, doi: 10.1177/07342829221141934.
[17] B. Zeng, M. Jeon, and H. Wen, "How does item wording affect participants' responses in Likert scale? Evidence from IRT analysis," *Frontiers in Psychology*, vol. 15, p. 1304870, Oct. 2024, doi: 10.3389/fpsyg.2024.1304870.
[18] H. Baumgartner, B. Weijters, and R. Pieters, "Misresponse to survey questions: a conceptual framework and empirical test of the effects of reversals, negations, and polar opposite core concepts," *Journal of Marketing Research*, vol. 55, no. 6, pp. 869–883, Dec. 2018, doi: 10.1177/0022243718811848.
[19] M. Lindwall *et al.*, "Method effects: the problem with negatively versus positively keyed items," *Journal of Personality Assessment*, vol. 94, no. 2, pp. 196–204, Mar. 2012, doi: 10.1080/00223891.2011.645936.
[20] J. Suárez-Alvarez, I. Pedrosa, L. M. Lozano, E. García-Cueto, M. Cuesta, and J. Muñiz, "Using reversed items in Likert scales: a questionable practice," *Psicothema*, vol. 30, no. 2, pp. 149–158, May 2018, doi: 10.7334/psicothema2018.33.
[21] N. A. Bowling, J. L. Huang, C. B. Bragg, S. Khazon, M. Liu, and C. E. Blackmore, "Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality," *Journal of Personality and Social Psychology*, vol. 111, no. 2, pp. 218–229, 2016, doi: 10.1037/pspp0000085.
[22] F. Antoniou and M. H. Alghamdi, "Confidence in mathematics is confounded by responses to reverse-coded items," *Frontiers in Psychology*, vol. 15, p. 1489054, Oct. 2024, doi: 10.3389/fpsyg.2024.1489054.
[23] F. A. Setiawati, S. R. Nurhayati, R. N. Amelia, and A. A. Darojat, "Study on the threats of reverse-worded items to the psychometric properties of the marital quality scale," *The Open Psychology Journal*, vol. 15, no. 1, pp. 1–8, Nov. 2022, doi: 10.2174/18743501-v15-e2208150.
[24] Y. J. Mudi, A. Kasrianti, S. P. B. Syahailatua, N. Isnaini, and B. Eba, "Negative-worded items functioning as method artifacts in the chemistry identity scale: evidence from exploratory, confirmatory, and bifactor analyses," *Integrated Science Education Journal*, vol. 7, no. 3, pp. 459–469, May 2026, doi: 10.37251/isej.v7i3.2960.
[25] S. Y. Y. Chyung, J. R. Barkin, and J. A. Shamsy, "Evidence-based survey design: the use of negatively worded items in surveys," *Performance Improvement*, vol. 57, no. 3, pp. 16–25, Mar. 2018, doi: 10.1002/pfi.21749.
[26] H. C. Bulut and O. Bulut, "Item wording effects in self-report measures and reading achievement: does removing careless respondents help?" *Studies in Educational Evaluation*, vol. 72, p. 101126, Mar. 2022, doi: 10.1016/j.stueduc.2022.101126.
[27] J. N. Sneddon, K. A. Adelaar, D. N. Djenar, and M. C. Ewing, *Indonesian: a comprehensive grammar*, 2nd ed. London, U.K.: Routledge, 2012.
[28] P. B. Smith, "Acquiescent response bias as an aspect of cultural communication style," *Journal of Cross-Cultural Psychology*, vol. 35, no. 1, pp. 50–61, Jan. 2004, doi: 10.1177/0022022103260380.
[29] J. He and F. van de Vijver, "Bias and equivalence in cross-cultural research," *Online Readings in Psychology and Culture*, vol. 2, no. 2, pp. 1–19, Jun. 2012, doi: 10.9707/2307-0919.1111.
[30] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate data analysis*, 8th ed. Andover, Hampshire, U.K.: Cengage Learning, 2019.
[31] R. B. Kline, *Principles and practice of structural equation modeling*, 4th ed. New York, NY: The Guilford Press, 2016.
[32] E. J. Wolf, K. M. Harrington, S. L. Clark, and M. W. Miller, "Sample size requirements for structural equation models: an evaluation of power, bias, and solution propriety," *Educational and Psychological Measurement*, vol. 73, no. 6, pp. 913–934, Dec. 2013, doi: 10.1177/0013164413495237.
[33] L. R. Aiken, "Content validity and reliability of single items or questionnaires," *Educational and Psychological Measurement*, vol. 40, no. 4, pp. 955–959, Dec. 1980, doi: 10.1177/001316448004000419.
[34] M. S. B. Yusoff, "ABC of content validation and content validity index calculation," *Education in Medicine Journal*, vol. 11, no. 2, pp. 49–54, Jun. 2019, doi: 10.21315/eimj2019.11.2.6.
[35] N. Kania, Y. S. Kusumah, J. A. Dahlan, E. Nurlaelah, F. Gürbüz, and E. Bonyah, "Constructing and providing content validity evidence through the Aiken's V index based on the experts' judgments of the instrument to measure mathematical problem-solving skills," *REID (Research and Evaluation in Education)*, vol. 10, no. 1, pp. 64–79, Jun. 2024, doi: 10.21831/reid.v10i1.71032.
[36] T. A. Brown, *Confirmatory factor analysis for applied research*, 2nd ed. New York, NY: The Guilford Press, 2015.
[37] M. Nu'man, H. Retnawati, Sugiman, and Jailani, "Measuring self-regulated learning in the STEM framework: a confirmatory factor analysis," *European Journal of Educational Research*, vol. 10, no. 4, pp. 2067–2077, Oct. 2021, doi: 10.12973/eu-jer.10.4.2067.
[38] B. M. Byrne, *Structural equation modeling with Mplus: basic concepts, applications, and programming*, 1st ed. New York, NY: Routledge, 2012.

[39] E. van Sonderen, R. Sanderman, and J. C. Coyne, "Ineffectiveness of reverse wording of questionnaire items: let's learn from cows in the rain," *PLoS ONE*, vol. 8, no. 7, p. e68967, Jul. 2013, doi: 10.1371/journal.pone.0068967.
[40] N. K. Malhotra, *Marketing research: an applied orientation*, 6th ed. Upper Saddle River, NJ: Pearson Education, 2010.
[41] D. Nunan, D. F. Birks, and N. K. Malhotra, *Marketing research: applied insight*, 6th ed. Harlow, U.K.: Pearson, 2020.
[42] D. George and P. Mallery, *IBM SPSS statistics 26 step by step: a simple guide and reference*, 16th ed. New York, NY: Routledge, 2020.
[43] B. Weijters and H. Baumgartner, "Misresponse to reversed and negated items in surveys: a review," *Journal of Marketing Research*, vol. 49, no. 5, pp. 737–747, Oct. 2012, doi: 10.1509/jmr.11.0368.
[44] J. W. Creswell and V. L. P. Clark, *Designing and conducting mixed methods research*, 3rd ed. Thousand Oaks, CA: SAGE Publications, Inc., 2018.
[45] O. R. Imawan, H. Retnawati, Haryanto, and R. Ismail, "Confirmatory factor analysis and differential item functioning analysis on mathematical literacy instruments for prospective Indonesian elementary school teachers," in *AIP Conference Proceedings*, 2024, p. 080009, doi: 10.1063/5.0228174.
[46] D. Apriandi, H. Retnawati, and A. M. Abadi, "Construct validity and reliability of the learning motivation questionnaire," *TEM Journal*, vol. 11, no. 4, pp. 1494–1499, Nov. 2022, doi: 10.18421/TEM114-09.
[47] A. W. Meade and S. B. Craig, "Identifying careless responses in survey data," *Psychological Methods*, vol. 17, no. 3, pp. 437–455, Sep. 2012, doi: 10.1037/a0028085.

BIOGRAPHIES OF AUTHORS



Davi Apriandi is an associate professor at the Faculty of Teacher Training and Education, Universitas PGRI Madiun, Indonesia. His research interests encompass mathematics education, the TPACK framework, computational thinking, instructional design, and the development of learning materials and research instruments. He can be contacted at email: davi.mathedu@unipma.ac.id.



Heri Retnawati is a professor in the field of Mathematics Education Evaluation and a lecturer at Universitas Negeri Yogyakarta, Indonesia. Her research focuses on mathematics education, evaluation of mathematics education, and the measurement and assessment of mathematics education. She can be contacted at email: heri_retnawati@uny.ac.id.



Agus Maman Abadi is a professor of Mathematics and a lecturer at Universitas Negeri Yogyakarta, Indonesia. His research focuses on fuzzy systems, neural networks, deep learning and its applications, and applied linear algebra. He can be contacted at email: agusmaman@uny.ac.id.



Annis Deshinta Ayuningtyas is an assistant professor at the Faculty of Teacher Training and Education, Universitas Sarjanawiyata Tamansiswa, Indonesia. Her research interests include mathematics education, learning evaluation, STEM-culture integration, and ethnomathematics. She can be contacted at email: annis.deshinta@ustjogja.ac.id.