

EVALUATING A LOCAL WISDOM-BASED RELIGIOUS MODERATION ATTITUDE INSTRUMENT USING RASCH MODELING

Mujiansyah¹, Hamdan², Tamjidnor³, Muhammad Nur Effendi⁴

¹²³⁴Universitas Islam Negeri Antasari Banjarmasin; Indonesia

Correspondence Email; mujiansyah@uin-antasari.ac.id

Submitted: 14/07/2026

Revised: 11/08/2026

Accepted: 09/09/2026

Published: 14/09/2026

Abstract

To address this gap, this study aims to evaluate the temporal stability and reliability of the Local Wisdom-Based Religious Moderation Attitude Questionnaire for Islamic Junior High School (Madrasah Tsanawiyah/MTs) students using the Rasch Modeling paradigm. This study employed a quantitative approach, specifically utilizing a cross-sectional survey design combined with a test-retest method over a three-week interval. Involving a cohort of 375 students from six MTs in South Kalimantan Province, the research evaluated a 16-item instrument utilizing a 4-point forced-choice Likert scale. Data were computationally analyzed using the *Rating Scale Model* (RSM) via Winsteps software to extract interval reliability metrics, temporal stability, scale diagnostics, unidimensionality, and item fit. The empirical analysis confirmed that the instrument fulfilled the unidimensionality assumption, with a raw variance explained of 48.5%, substantiating that the dimensions integrate into a single construct. The test-retest assessment demonstrated robust temporal stability, evidenced by the logit correlation of the Person Measure (0.89) and Item Measure (0.94), indicating the instrument's resilience against temporary cognitive fluctuations. Furthermore, the internal reliability proved to be excellent (Cronbach's Alpha 0.88; Person Reliability 0.86; Item Reliability 0.95), complemented by a person separation index (2.54) that accurately differentiates students' affective strata. Scale diagnostics verified that the exclusion of a neutral option allowed the instrument to operate hierarchically without threshold disordering, and all 16 items were proven to fit the model. In conclusion, this local wisdom-based moderation instrument exhibits exceptional logical consistency and temporal stability. Therefore, it is highly recommended as an early warning system and a standard for affective curriculum evaluation for educational policymakers.

Keywords

Reliability Evaluation; Religious Moderation; Local Wisdom; Rasch Modeling; Test-Retest Stability.



© 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY NC) license (<https://creativecommons.org/licenses/by-nc/4.0/>).

INTRODUCTION

Sociocultural diversity in Indonesia offers rich social capital, yet it simultaneously creates exposure to ideological polarization and religious exclusivism. Empirical surveys consistently report rising intolerance among secondary school students, raising concerns for social cohesion (Mukhibat et al., 2023; Nur Adnan Saputra et al., 2021). Strengthening religious moderation (*wasathiyyah*) in schools is therefore a priority. Islamic Junior High Schools (Madrasah Tsanawiyah/MTs), administered under the Ministry of Religious Affairs, occupy a crucial position in this effort. Because MTs adolescents undergo rapid cognitive and identity development, early exposure to inclusive values serves as an effective safeguard against radicalization. Religious moderation cannot be evaluated without considering local community realities. The Ministry of Religious Affairs identifies four core indicators of moderation: national commitment, tolerance, non-violence, and accommodation of local customs. Integrating regional cultural wisdom distinguishes Indonesian *wasathiyyah* practice from broader theological constructs (Putri & Fadlullah, 2022). Local traditions act as cultural anchors, making abstract religious concepts tangible and accessible to students (Ani, 2025).

Sociological evidence shows that indigenous customs, cooperative community norms, and cultural heritage buffer communities against exclusive ideologies (Putra et al., 2021). Consequently, affective evaluation instruments must explicitly embed these regional cultural dimensions alongside standard moderation indicators. Despite growing interest in religious moderation, existing assessment practices in Islamic education exhibit clear psychometric shortcomings. Previous inquiries broadly fall into two groups. The first group examines pedagogical methods and general student attitudes using qualitative observations or basic descriptive surveys (Muaz & Ruswandi, 2022; Santoso et al., 2022; Subchi et al., 2022). While these studies document qualitative trends, they do not produce standardized, calibrated measurement scales. The second group focuses on affective scale development, yet predominantly relies on Classical Test Theory (CTT) (Sumintono & Widhiarso, 2014). CTT models present known limitations: reliability estimates depend heavily on sample characteristics, and ordinal Likert scores are improperly analyzed as interval data (Bintang & Suprananto, 2024; Boone et al., 2014). Modern psychometrics addresses these limitations through Item Response Theory (IRT) and Rasch modeling, which convert ordinal responses into linear logit measures (Cook & Wind, 2024).

In addition, forced-choice rating scales within Rasch models remove central tendency bias and yield sample-invariant parameters (Combrinck, 2024). Nevertheless, researchers have yet to combine a test-retest logit calibration design with a local wisdom-based moderation scale for secondary students. To bridge this gap, this study evaluates the psychometric reliability, temporal logit stability, and scale functioning of the Local Wisdom-Based Religious Moderation Attitude Questionnaire across MTs students in South Kalimantan using Rasch analysis. Four specific objectives guide this investigation: 1. To estimate internal consistency, item/person reliability, and person separation indices. 2. To examine test-retest stability for person and item measures across a three-week interval on an invariant logit scale. 3. To evaluate the category functioning of the 4-point forced-choice Likert scale and check for threshold disordering. 4. To verify unidimensionality and item fit expectations, confirming that the scale captures a single latent construct with minimal measurement error. By establishing these psychometric parameters, this work provides a reliable, calibrated evaluation tool for guidance counselors, educators, and educational policy planners in madrasahs.

In the specific research context of South Kalimantan, Islamic Junior High Schools (Madrasah Tsanawiyah/MTs) have actively implemented institutional directives to foster religious moderation (*wasathiyyah*), integrating principles of tolerance, non-violence, and social cohesion into subjects such as *Akidah Akhlak* and co-curricular madrasah activities (El Hasbi & Noor Fuady, 2024; Sabri et al., 2022). These pedagogical efforts are historically supported by South Kalimantan's rich sociocultural heritage, where classic local Islamic scholar traditions—such as Syekh Muhammad Arsyad al-Banjari's *Sabilal Muhtadin*—and communal values like *Kayuh Baimbai* (cooperative solidarity) serve as cultural anchors for religious harmony (Arif, 2020; Syaifullah, 2020). However, empirical observations across madrasahs in South Kalimantan reveal a complex and non-uniform picture regarding students' actual perceptions of religious moderation. While students generally demonstrate strong formal adherence to national commitment, empirical surveys indicate lingering attitudinal variance when navigating complex issues of theological pluralism, cross-cultural traditions, and exposure to digital exclusivity (Jubba et al., 2021; Lestari, 2020). Furthermore, co-curricular moderation programs frequently operate without systematically evaluating how students internalize these values, leading to potential discrepancies between official policy goals and the genuine affective stances of junior high school learners (Nabiilah & Arifin, 2026; Nazaruddin Iqbal Al Hamiid & Madjid, 2025).

This empirical divergence highlights a critical academic concern: assessing affective constructs like religious moderation among early adolescents (ages 12–15) is inherently susceptible to developmental cognitive instability, peer influence, and temporary response fluctuations (Agusfianto et al., 2023; Hayatunnisa et al., 2024). In educational settings, uncalibrated evaluation instruments frequently capture transient opinion shifts or social desirability bias rather than stable, enduring core attitudes (Arif, 2020; Septikasari et al., 2024; Wahyudi et al., 2022). Consequently, evaluating the temporal stability and measurement precision of a local wisdom-based moderation questionnaire across time intervals is an absolute empirical prerequisite before such tools can be reliably utilized for affective curriculum evaluation or diagnostic policy planning (Anwar, 2021; Zakariyah et al., 2022). Relying on single-administration surveys without establishing invariant item-person parameters risks generating misleading diagnostic data regarding student radicalization risk or tolerance levels. Establishing a psychometrically rigorous, test-retest calibrated instrument grounded in modern Item Response Theory thus becomes essential to guarantee that observed scores accurately reflect stable latent traits rather than measurement noise within pluralistic secondary education environments (Fitri et al., 2018; Wyse, 2010).

The second stream of literature addresses affective scale measurement in educational contexts. Traditionally, evaluation tools for religious attitudes in Indonesia have relied heavily on Classical Test Theory (CTT) (Sumintono & Widhiarso, 2014). However, as demonstrated by Bintang, Suprananto, and Boone, CTT approaches suffer from intrinsic flaws: reliability indices are highly sample-dependent, and ordinal Likert responses are frequently misinterpreted as equal-interval data (Bintang & Suprananto, 2024; Boone et al., 2014). To resolve these measurement distortions, recent psychometric literature strongly advocates for Item Response Theory (IRT) and Rasch Modeling (Engelhard & Wind, 2017). Furthermore, Combrinck demonstrated that forced-choice Rasch rating scales effectively eliminate central tendency bias and yield invariant logit parameters (Combrinck, 2024). Despite these methodological advancements, prior studies have not applied the Rasch Modeling paradigm—specifically incorporating a test-retest logit calibration design—to evaluate a local wisdom-integrated religious moderation instrument for secondary school students. Positioned at the intersection of culturally grounded affective construct design and modern IRT psychometrics, this study fills this critical empirical gap.

To address this gap, this study systematically aims to evaluate the psychometric reliability, temporal stability, and scale functioning of the Local Wisdom-Based Religious Moderation Attitude

Questionnaire for Islamic Junior High School (MTs) students using the Rasch Modeling paradigm. Specifically, this study addresses four operational objectives: 1. To evaluate the internal consistency, item reliability, person reliability, and separation indices of the instrument. 2. To examine the temporal stability (test-retest reliability) of person and item measures across a three-week interval on an invariant logit scale. 3. To diagnose the functionality of the 4-point forced-choice Likert rating scale and verify the absence of category threshold disordering, and to verify the psychometric prerequisites of unidimensionality and item fit, ensuring that the instrument accurately measures a unified latent construct with minimal measurement error. The empirical evidence generated by this investigation aims to establish a standardized, highly stable evaluation reference for educators and strategic policymakers in madrasahs.

METHOD

This study employed a quantitative approach, combining a cross-sectional survey design with a test-retest method to evaluate the psychometric reliability of the Local Wisdom-Based Religious Moderation Attitude instrument. The primary objective of this design was to diagnose response inconsistencies, minimize measurement errors, and examine the temporal stability of the scale across time (Cohen et al., 2007; Creswell & Plano Clark, 2023). Instrument reliability testing was executed using the Rasch Modeling paradigm, specifically under Item Response Theory (IRT), which is globally recognized for its ability to independently isolate person parameters and item difficulty levels, thereby generating precise interval logit reliability estimates (Stemler & Naples, 2021; Vijayamohanan Pillai N & Rjumohan Asalatha, 2020).

The study was conducted in South Kalimantan Province between April 2 and May 23, 2026, comprising two distinct data collection phases with a three-week interval to fulfill the prerequisites for test-retest stability assessment. This region was strategically selected due to its rich sociocultural landscape, which effectively grounds the local wisdom variables within religious moderation. The three-week time interval was specifically calibrated to prevent memory recall effects while ensuring the absence of drastic cognitive or developmental fluctuations regarding affective moderation values during the evaluation period (Gelino et al., 2024; Hossan et al., 2025). The target population comprised Islamic Junior High School (Madrasah Tsanawiyah/MTs) students in South Kalimantan, who are developmentally in an identity-formation phase susceptible to radical exposure. Utilizing a combined purposive and convenience sampling strategy, a final sample of 375 complete and analyzable student datasets was successfully gathered across six participating madrasahs. In Rasch

calibration, this sample size ($N = 375$) is highly robust and vastly exceeds minimum requirements, mathematically guaranteeing stable item calibration parameters with a high degree of replicability across broader populations (Tutz, 2024; Zijlmans et al., 2019).

To provide a comprehensive evaluation, this study systematically utilized both primary and secondary data sources: Primary data consisted of raw item responses and calibrated logit measures obtained directly from the 375 MTs student participants across two data collection cycles (Phase 1/Test and Phase 2/Retest). The primary data captured students' self-reported affective scores regarding national commitment, tolerance, anti-violence, acceptance of tradition, and local wisdom. Secondary data were gathered to establish the institutional context, sample frame verification, and curriculum mapping. These sources included official demographic and administrative statistics from the Regional Office of the Ministry of Religious Affairs (Kanwil Kemenag) South Kalimantan, student enrollment rosters from participating madrasahs, and official policy guidelines regarding religious moderation educational frameworks.

The evaluated instrument comprised 16 statement items covering five sub-dimensions: National Commitment, Tolerance, Anti-Violence, Acceptance of Tradition, and Local Wisdom. The rating scale utilized a 4-point forced-choice Likert scale (1 = Strongly Disagree, 2 = Disagree, 3 = Agree, 4 = Strongly Agree). The omission of a neutral or middle category was intentional—it eliminates central tendency bias, forces respondents to take a definitive stance, and significantly reduces the Standard Error of Measurement (SEM) in adolescent psychometric inventories (Ekstrand et al., 2022; Nie et al., 2024).

Framed within the Rasch psychometric modeling framework, this study tested four operational hypotheses: H_{a1} (Internal Consistency and Separation): The instrument exhibits high internal consistency, with both Person Reliability (R_{person}) and Item Reliability (R_{item}) exceeding 0.80, alongside a Person Separation Index > 2.0 (Bond et al., 2020). H_{a2} (Temporal Stability): The instrument demonstrates strong temporal stability between Phase 1 and Phase 2, evidenced by a statistically significant logit correlation ($r \geq 0.80, p < 0.05$) for both Person Measures and Item Measures across the three-week interval. H_{a3} (Scale Rating Functionality): The 4-point forced-choice Likert scale operates monotonically without threshold disordering (ordered Andrich Thresholds) (Linacre, 2002), and H_{a4} (Construct Validity and Model Fit): The instrument satisfies the unidimensionality requirement (Raw Variance Explained $> 20\%$) and all 16 items fit the Rasch Rating Scale Model ($0.5 \leq \text{Outfit MNSQ} \leq 1.5$) (Boone et al., 2014; Trizano-Hermosilla et al., 2021).

Raw response data from both testing cycles were computationally analyzed using Winsteps software (Version 5.4.0) via the Rating Scale Model (RSM). The data analysis prioritized the following evaluation metrics: (1) Person and Item Reliability indices (> 0.80 threshold); (2) Separation Index to evaluate respondent stratification; (3) Test-retest Pearson/Spearman logit correlations; (4) Rating scale category diagnostic thresholds; and (5) Unidimensionality and Item Fit statistics (Outfit MNSQ and Point-Measure Correlation) to ensure the scale accurately captures a single latent construct without measurement noise (Malhotra, 1987; Trizano-Hermosilla et al., 2021).

FINDINGS AND DISCUSSION

Findings

This study analyzed the responses of 375 Islamic Junior High School (Madrasah Tsanawiyah/MTs) students in South Kalimantan Province to the 16 statement items of the Local Wisdom-Based Religious Moderation Attitude Questionnaire. Given that the primary objective of this research centered on reliability evaluation, the data analysis—based on Rasch Modeling (Rating Scale Model) using Winsteps software—was profoundly focused on the fundamental indicators of interval stability and logical response consistency.

The analysis results are presented across four main metrics: reliability and separation levels, test-retest correlation, rating scale diagnostics, and the fulfillment of unidimensionality and item fit prerequisites.

Reliability and Separation Levels

The most crucial step in this evaluation is assessing the internal consistency of the instrument's operation. Table 1 summarizes the instrument's reliability parameters.

Table 1. Summary of Reliability and Separation Statistics

Measurement Parameter	Reliability	Value	Separation Index	Quality Interpretation
<i>Person (Respondent Stability)</i>	0,86		2,54	Good
<i>Item (Item Stability)</i>	0,95		4,36	Excellent
<i>Cronbach's Alpha (KR-20)</i>	0,88		-	Excellent

Referring to Table 1, the instrument possesses a Cronbach's Alpha value of 0.88, indicating that the overall internal reliability falls into the excellent category. Through a more specific metric, the Item Reliability reached 0.95, demonstrating that the quality of item difficulty levels is exceptionally defined and stable. Furthermore, the Person Separation value of 2.54 proves the instrument's precision in accurately categorizing students into three distinct affective ability strata.

Temporal Stability (Test-Retest Correlation)

To minimize measurement errors, a temporal stability analysis was conducted based on two data collection phases with a three-week interval.

Table 2. Test-Retest Correlation Analysis (Logit Scale Stability)

Correlation Parameter	Correlation Coefficient (<i>r</i>)	Significance value)	Stability Criterion
Person Measure Correlation (Phase 1 vs. Phase 2)	0,89	< 0,01	Very Strong / Stable
Item Measure Correlation (Phase 1 vs. Phase 2)	0,94	< 0,01	Very Strong / Stable

The test-retest correlation results presented in Table 2 empirically verify the absence of drastic fluctuations in student response consistency throughout the testing interval. Furthermore, the logit scale proved highly robust in withstanding distortion effects.

Rating Scale Diagnostics

The reliability of the instrument is also significantly determined by the students' optimal comprehension of the response options. This questionnaire adopted a 4-point forced-choice Likert scale to compel subjects to express a definitive stance.

Table 3. Rating Scale Diagnostics (4 Categories)

Response Category	Score	Response Count	Observed Count (%)	Average Measure (Logit)	Andrich Threshold (Logit)
Strongly Disagree (SD)	1	450	7,5%	-1,85	NONE
Disagree (D)	2	1.125	18,7%	-0,62	-1,48
Agree (A)	3	2.580	43,0%	0,84	0,25
Strongly Agree (SA)	4	1.845	30,8%	1,95	1,23

Table 3 demonstrates that the Average Measure values increase monotonically from the lowest to the highest category. The Andrich Threshold values also advance sequentially (-1.48 -> 0.25 -> 1.23), confirming the absence of threshold disordering. Consequently, the scale hierarchy functions perfectly in capturing the gradations of the respondents' attitudes.

Fulfillment of Unidimensionality and Item Fit Prerequisites

As an absolute prerequisite to ensure that the accuracy of the final reliability estimates does not distort the measurement, the instrument must be proven to measure only a single dimension.

Table 4. Item Fit Order

Item Code	Indicator	Measure (Logit)	Outfit MNSQ	Outfit ZSTD	Pt-Measure Corr	Status
Item_16	Local Wisdom	1.25	1.05	0.4	0.65	FIT
Item_14	Local Wisdom	0.98	0.88	-1.1	0.72	FIT
Item_05	Tolerance	0.85	1.12	1.2	0.58	FIT
Item_08	Anti-Violence	0.76	1.35	1.9	0.45	FIT
Item_11	Acceptance of Tradition	0.65	0.95	-0.4	0.68	FIT
Item_02	National Commitment	0.55	0.82	-1.5	0.74	FIT
Item_15	Local Wisdom	0.42	1.01	0.1	0.66	FIT
Item_09	Anti-Violence	0.25	0.94	-0.6	0.70	FIT
Item_12	Acceptance of Tradition	0.10	1.15	1.4	0.55	FIT
Item_04	Tolerance	-0.15	0.85	-1.2	0.71	FIT
Item_01	National Commitment	-0.35	1.06	0.5	0.64	FIT
Item_13	Acceptance of Tradition	-0.45	1.22	1.7	0.50	FIT
Item_06	Tolerance	-0.65	0.78	-1.8	0.75	FIT
Item_10	Anti-Violence	-0.95	0.91	-0.8	0.69	FIT
Item_07	Tolerance	-1.35	1.08	0.7	0.62	FIT
Item_03	National Commitment	-1.90	0.96	-0.3	0.67	FIT

Note: Fit Criteria = Outfit MNSQ 0.5 - 1.5; Outfit ZSTD -2.0 to +2.0; Pt-Measure Correlation positive (0.4 - 0.8).

Table 5. Summary of Rasch Measurement Prerequisites

Prerequisite Evaluation Criteria			Observed Value	Rasch Acceptance Criteria	Fulfillment Status
Unidimensionality: Raw variance explained			48,5%	> 20% (Ideal)	Fulfilled (Cohesive Construct)
Unidimensionality: Unexplained variance in 1st contrast			6,5%	< 15% (Ideal)	Fulfilled
Item Fit: Outfit MNSQ Range (16 Items)			0,78 s.d. 1,35	0,5 s.d. 1,5	All 16 Items FIT
Item Fit: Pt-Measure Corr Range (16 Items)			0,45 s.d. 0,75	Positive Value (>0,40)	All 16 Items FIT

Based on Tables 4 and 5, the raw variance explained reached 48.5%, proving that the five sub-dimensions are strongly integrated into a single latent construct. Furthermore, all *Outfit MNSQ* values for the 16 statement items fell within the ideal range (0.5 to 1.5), exhibiting positive point-measure correlations. Since all items fulfilled the fit prerequisites and exhibited no ambiguity, no questionnaire items were eliminated from this reliability evaluation.

Discussion

This study aims to evaluate the reliability of the Local Wisdom-Based Religious Moderation Attitude instrument for Islamic Junior High School (Madrasah Tsanawiyah/MTs) students using Rasch Modeling. The selection of the Rasch Model has proven to provide a significantly superior, precise, and objective psychometric evaluation compared to Classical Test Theory (CTT), which is susceptible to recurrent measurement errors in stability testing (Cook & Wind, 2024; Stemler & Naples, 2021). This interval reliability evaluation empirically confirms the instrument's feasibility for implementation as a stable diagnostic parameter.

The most crucial finding of this research is the proven temporal stability of the instrument through the test-retest assessment. The cross-temporal logit correlation results (Person Measure 0.89 and Item Measure 0.94) confirm that this instrument is immune to temporary cognitive fluctuations. This logit score consistency is highly vital in affective assessments, where the vulnerability of adolescent subjects to shifts in opinion often fatally undermines scale stability (Struijs et al., 2020; Uberti, 2022). This empirical stability proves that the measured indicators of wasathiyyah and local wisdom are deeply rooted as traits (solid characters) rather than mere states (temporary conditions) within the memory of the research subjects (Castro-Alvarez et al., 2022).

The internal reliability level and separation index also demonstrate highly robust operational performance. The instrument recorded an Item Reliability of 0.95 and a Person Separation of 2.54. These parameters are not merely superficial statistical figures; rather, they serve as evidence that the item calibration is highly precise in stratifying respondents' abilities into three determinant strata: highly moderate, fairly moderate, and vulnerable to radicalism. This sharpness in diagnostic stratification guarantees that the instrument is free from floor or ceiling effects, an absolute prerequisite for affective measurement tools mapping the risk profile of exclusivism in madrasahs (Engelhard & Wang, 2021; Magazzù et al., 2022).

The superior reliability of this calibration is evidently significantly influenced by the design of the 4-point forced-choice Likert rating scale function. The absence of a neutral option effectively stimulates students to engage in deep cognitive processing before responding, rather than taking refuge behind a central tendency bias that obscures data consistency. The confirmation of the Andrich Threshold progressing hierarchically without threshold disordering proves that the cognitive load of these even-numbered options is optimally resolved by the MTs student demographic. This middle-option elimination approach has been universally proven capable of

reducing the Standard Error of Measurement (SEM) and multiplying logit accuracy within the adolescent age range (Combrinck, 2024; Ekstrand et al., 2022).

The high estimation of the instrument's reliability is absolutely validated by the fulfillment of the unidimensionality prerequisite, reaching 48.5% raw variance explained. Meeting this foundational assumption secures the claim that the strong test-retest correlation genuinely stems from the measurement of a cohesive dimension, rather than a statistical artifact (Weir, 2005). This also reinforces the theoretical paradigm that the local wisdom dimension is a cohesive formative variable with the wasathiyyah attitude, ensuring that its simultaneous measurement does not distort the instrument's reliability parameters (Raykov & Zhang, 2024; Trizano-Hermosilla et al., 2021).

All 16 statement items were proven to fit the model parameters; thus, none were eliminated during this interval reliability validation process. The proportional Outfit MNSQ range eliminates potential anomalies of measurement noise caused by random guessing or repetitive responses. The instrument's cleanliness from response data distortion serves as a strong indication that the instrument's phrasing—such as the item on non-violent conflict mediation—aligns with the psychosocial capacity and cultural integrity at the secondary education level (Blackwell et al., 2017; Ridwan & Mahmudi, 2023; Tayfur et al., 2021).

Definitively, this research has successfully transformed a conceptual questionnaire design into a diagnostic instrument unit possessing exceptional temporal stability and logical consistency. This instrument does not merely fill the gap in religious evaluation measurement literature; it establishes a standard baseline parameter. Guidance counselors and policymakers are now equipped with a reliable early warning system protocol, which is essential for monitoring the evaluative outcomes of the moderation curriculum year over year without compromising data accuracy (Jin et al., 2025; Munawar et al., 2024).

CONCLUSION

Based on the empirical findings, the Local Wisdom-Based Religious Moderation Attitude Questionnaire demonstrates exceptional interval reliability, logical consistency, and temporal stability when evaluated through the Rasch Modeling paradigm among Islamic Junior High School (MTs) students. The test-retest logit correlations confirm the instrument's resilience against temporary cognitive fluctuations across a three-week interval (Person Measure = 0.89; Item Measure = 0.94), supported by robust internal reliability parameters (Item Reliability = 0.95; Person Separation = 2.54) and monotonic 4-point forced-choice rating scale performance. Consequently, this calibrated

questionnaire offers educational policymakers and counselors a highly objective, distortion-free diagnostic tool for tracking affective moderation traits. Nevertheless, certain empirical boundaries must be acknowledged. Because the sample ($N = 375$) was drawn exclusively from madrasahs in South Kalimantan, the psychometric parameters may face generalizability constraints when applied across different regional cultural ecosystems or non-Islamic general secondary education settings. Additionally, despite the three-week test-retest interval, reliance on a self-report survey format leaves the instrument somewhat susceptible to social desirability bias among early adolescents.

To address these empirical constraints, future research should pursue multi-center cross-regional replication studies across Indonesia to verify measurement invariance across diverse sociocultural zones and among heterogeneous student populations in public and private secondary schools. Furthermore, researchers are encouraged to adopt mixed-methods frameworks that pair Rasch-calibrated logit scores with direct behavioral observations, thereby validating whether calibrated survey responses directly correspond to students' real-world manifestations of tolerance and social cohesion in the field. Ultimately, by establishing a psychometrically sound foundation that bridges regional cultural heritage with modern Item Response Theory, this study provides a stable, standardized baseline for evaluating affective curriculum outcomes and informing targeted educational policies in pluralistic secondary education contexts.

REFERENCES

- Agusfianto, N. P., Andayani, S. U., Anwar, Sihombing, L., Setiawan, A., Amelia, D., Gunawan, C., Indraswati, T. D., Wati, K. M., Octasyilva, A. R. P., Sari, D. P., Nurbismi, Kamal, E., Luthfiyani, U. K., Sijabat, F. N., & Wijayanto, S. A. (2023). *Pengantar Bisnis (Respons Dinamika Era Digital)*. Thesis Commons. <https://doi.org/10.31237/osf.io/mrzs5>
- Aluf, W. A., Bukhori, I., & Bashith, A. (2024). Evaluasi Pembelajaran Moderasi Beragama untuk Mengukur Penguatan Toleransi Siswa di MIN 2 Pamekasan. *Jurnal Pendidikan dan Pembelajaran Indonesia (JPPI)*, 4(4), 1623–1634. <https://doi.org/10.53299/jppi.v4i4.825>
- Ani, A. (2025). Peran Kearifan Lokal sebagai Mediator Hubungan Moderasi Beragama dalam Hubungan Moderasi Beragama terhadap Kerukunan Umat Islam di Indonesia. *At-Thullab : Jurnal Mahasiswa Studi Islam*, 7(2), 232–250. <https://doi.org/10.20885/tullab.vol7.iss2.art2>
- Anwar, K. (2021). Pancasila Village, Multicultural Education and Moderation of Diversity in Indonesia. *Nazhruna: Jurnal Pendidikan Islam*, 4(2), 221–234. <https://doi.org/10.31538/nzh.v4i2.1238>
- Arif, A. M. (2020). Perspektif Teori Sosial Emile Durkheim dalam Sosiologi Pendidikan. *Moderasi: Jurnal Studi Ilmu Pengetahuan Sosial*, 1(2), 1–14. <https://doi.org/10.24239/moderasi.Vol1.Iss2.28>
- Bintang, R. S., & Suprananto, S. (2024). The Impact of Sample Size, Test Length, and Person-Item Targeting on the Separation Reliability in the Rasch Model: A Simulation Study. *JP3I (Jurnal Pengukuran Psikologi dan Pendidikan Indonesia)*, 13(1), 1–13.

- <https://doi.org/10.15408/jp3i.v13i1.27975>
- Blackwell, M., Honaker, J., & King, G. (2017). A Unified Approach to Measurement Error and Missing Data: Overview and Applications. *Sociological Methods & Research*, 46(3), 303–341. <https://doi.org/10.1177/0049124115585360>
- Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. Springer Netherlands. <https://doi.org/10.1007/978-94-007-6857-4>
- Castro-Alvarez, S., Tendeiro, J. N., Meijer, R. R., & Bringmann, L. F. (2022). Using Structural Equation Modeling to Study Traits and States in Intensive Longitudinal Data. *Psychological Methods*, 27(1), 17–43. <https://doi.org/10.1037/met0000393>
- Cohen, L., Manion, L., & Morrison, K. (2007). *Research Methods in Education* (0 ed.). Routledge. <https://doi.org/10.4324/9780203029053>
- Combrinck, C. (2024). Not Liking the Likert? A Rasch Analysis of Forced-choice Format and Usefulness in Survey Design. *Sage Open*, 14(4), 21582440241295501. <https://doi.org/10.1177/21582440241295501>
- Cook, R. M., & Wind, S. A. (2024a). Item Response Theory: A Modern Measurement Approach to Reliability and Precision for Counseling Researchers. *Measurement and Evaluation in Counseling and Development*, 57(2), 116–135. <https://doi.org/10.1080/07481756.2023.2301284>
- Cook, R. M., & Wind, S. A. (2024b). Item Response Theory: A Modern Measurement Approach to Reliability and Precision for Counseling Researchers. *Measurement and Evaluation in Counseling and Development*, 57(2), 116–135. <https://doi.org/10.1080/07481756.2023.2301284>
- Creswell, J. W., & Plano Clark, V. L. (2023). Revisiting Mixed Methods Research Designs Twenty Years Later. Dalam C. Poth (Ed.), *The Sage Handbook of Mixed Methods Research Design* (hlm. 21–36). Sage Publications Ltd. <https://doi.org/10.4135/9781529614572.n6>
- Ekstrand, J., Westergren, A., Årestedt, K., Hellström, A., & Hagell, P. (2022). Transformation of Rasch Model Logits for Enhanced Interpretability. *BMC Medical Research Methodology*, 22(1), 332. <https://doi.org/10.1186/s12874-022-01816-1>
- El Hasbi, A. Z. & Noor Fuady. (2024). Moderasi Beragama, Tasamuh, dan Sinkretisme (Dinamika Sosial Keagamaan Umat Islam). *Kamaliyah: Jurnal Pendidikan Agama Islam*, 2(1), 169–182. <https://doi.org/10.69698/jpai.v2i1.519>
- Engelhard, G., & Wang, J. (2021). *Rasch Models for Solving Measurement Problems: Invariant Measurement in the Social Sciences*. Sage. <https://doi.org/10.4135/9781071878675>
- Engelhard, G., & Wind, S. A. (2017). *Invariant Measurement with Raters and Rating Scales: Rasch Models for Rater-Mediated Assessments* (1 ed.). Routledge. <https://doi.org/10.4324/9781315766829>
- Fitri, E., Zola, N., & Ildil, I. (2018). Profil Kepercayaan Diri Remaja serta Faktor-Faktor yang Mempengaruhi. *JPPi (Jurnal Penelitian Pendidikan Indonesia)*, 4(1), 1–5. <https://doi.org/10.29210/02017182>
- Gelino, B. W., Schlitzer, R. D., Reed, D. D., & Strickland, J. C. (2024). A Systematic Review and Meta-Analysis of Test–Retest Reliability and Stability of Delay and Probability Discounting. *Journal of the Experimental Analysis of Behavior*, 121(3), 358–372. <https://doi.org/10.1002/jeab.910>
- Hayatunnisa Hayatunnisa, Jenika Fejrin, Milki Salwa Nor Azizah, Muhamad Ilham, Wayan Gastiadirrijal, Syahidin Syahidin, & Muhamad Parhan. (2024). Konsep Etika dan Moralitas Sebagai Materi dalam Pembelajaran Pendidikan Agama Islam. *Alfihris: Jurnal Inspirasi Pendidikan*, 2(2), 77–84. <https://doi.org/10.59246/alfihris.v2i2.765>
- Hossan, D., Wolfs, B., Jesmin, N.-E.-M., & Petkovic, M. (2025). *Questionnaire Validity and Reliability: A Review with Practical Guidelines*. 13, 135–186.
- Jin, R., Peng, Y., Wang, Z., Wang, J., Tang, J., & Zhang, M. (2025). Data-Driven Educational Decision-

- Making: How to Enhance Educational Quality and Management Efficiency. *Journal of Higher Education Research*, 5(6), 550. <https://doi.org/10.32629/jher.v5i6.3385>
- Jubba, H., Awang, J., & Sungkilang, S. A. (2021). The Challenges of Islamic Organizations in Promoting Moderation in Indonesia. *Wawasan: Jurnal Ilmiah Agama dan Sosial Budaya*, 6(1), 43–54. <https://doi.org/10.15575/jw.v6i1.12948>
- Lestari, J. (2020). Pluralisme Agama di Indonesia: Tantangan dan Peluang Bagi Keutuhan Bangsa. *Al-Adyan: Journal of Religious Studies*, 1(1), 29–38. <https://doi.org/10.15548/al-adyan.v1i1.1714>
- Linacre, J. (2002). *A User's Guide to Winsteps: Rasch-Model Computer Program*.
- Magazzù, G., Zampieri, G., & Angione, C. (2022). Clinical Stratification Improves the Diagnostic Accuracy of Small Omics Datasets within Machine Learning and Genome-Scale Metabolic Modeling Methods. *Computers in Biology and Medicine*, 151, 106244. <https://doi.org/10.1016/j.compbio.2022.106244>
- Malhotra, N. K. (1987). Validity and Structural Reliability of Multidimensional Scaling. *Journal of Marketing Research*, 24(2), 164–173. <https://doi.org/10.2307/3151506>
- Muaz, M., & Ruswandi, U. (2022). Moderasi Beragama dalam Pendidikan Islam. *JIIIP - Jurnal Ilmiah Ilmu Pendidikan*, 5(8), 3194–3203. <https://doi.org/10.54371/jiip.v5i8.820>
- Mukhibat, M., Nurhidayati Istiqomah, A., & Hidayah, N. (2023). Pendidikan Moderasi Beragama di Indonesia (Wacana dan Kebijakan). *Southeast Asian Journal of Islamic Education Management*, 4(1), 73–88. <https://doi.org/10.21154/sajiem.v4i1.133>
- Munawar, M., Aceng Kosasih, & Agus Fakhruddin. (2024). Implementasi Nilai-Nilai Moderasi Beragama dalam Pembelajaran PAI di Sekolah sebagai Upaya Pembentukan Karakter Moderat. *Didaktika: Jurnal Kependidikan*, 13(3 Agustus), 3413–3428. <https://doi.org/10.58230/27454312.848>
- Nabiilah, Z., & Arifin, N. R. (2026). Deep Learning Strengthens Religious Moderation Attitudes in Madrasah Students. *Indonesian Journal of Education Methods Development*, 21(2). <https://doi.org/10.21070/ijemd.v21i2.1010>
- Nazaruddin Iqbal Al Hamiid, & Madjid, Abd. (2025). A Systematic Literature Review on Strategies and Challenges in Instilling the Values of Religious Moderation at the Junior High School Level. *Lentera Pendidikan: Jurnal Ilmu Tarbiyah dan Keguruan*, 28(2), 466–483. <https://doi.org/10.24252/lp.2025v28n2i12>
- Nie, L., Xu, P., & Hu, D. (2024). Multidimensional IRT for Forced Choice Tests: A Literature Review. *Heliyon*, 10(5), e26884. <https://doi.org/10.1016/j.heliyon.2024.e26884>
- Nur Adnan Saputra, M., Nurul Mubin, M., Minhajul Abrori, A., & Handayani, R. (2021). Deradikalisasi Paham Radikal di Indonesia: Penguatan Kurikulum Pendidikan Islam Berbasis Moderasi. *Jurnal Pendidikan Agama Islam Al-Thariqah*, 6(2), 282–296. [https://doi.org/10.25299/al-thariqah.2021.vol6\(2\).6109](https://doi.org/10.25299/al-thariqah.2021.vol6(2).6109)
- Putra, A., Homsatun, A., Jamhari, J., Setiani, M., & Nurhidayah, N. (2021). Pemikiran Islam Wasathiyah Azyumardi Azra sebagai Jalan Moderasi Beragama. *Jurnal Riset Agama*, 1(3), 589–599. <https://doi.org/10.15575/jra.v1i3.15224>
- Putri, S. N. A., & Fadlullah, M. E. (2022). Wasathiyah Moderasi Moderasi Beragama) dalam Perspektif Quraish Shihab. *INCARE, International Journal of Educational Resources*, 3(1), 066–080. <https://doi.org/10.59689/incare.v3i1.390>
- Raykov, T., & Zhang, B. (2024). Evaluation of Maximal Reliability for Multidimensional Measuring Instruments Using Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 31(6), 1056–1064. <https://doi.org/10.1080/10705511.2024.2324458>
- Ridwan, A., & Mahmudi, A. F. (2023). The Internalization of Wasathiyah (Moderation) Values

- Through Pesantren-Based Education. *Journal of Educational Research and Practice*, 1(1), 60–72. <https://doi.org/10.70376/jerp.v1i1.30>
- Sabri, A., Nur Nst, W., & Halimah, S. (2022). Implementing Religious Moderation and Tolerance Attitude in Akidah Akhlak Learning. *Nazhruna: Jurnal Pendidikan Islam*, 5(3), 1266–1277. <https://doi.org/10.31538/nzh.v5i3.2649>
- Santoso, J., Saron, T. B., Sutrisno, S., & Putrawan, B. K. (2022). Moderasi Beragama di Indonesia: Kajian Tentang Toleransi dan Pluralitas di Indonesia. *Jurnal Teologi Berita Hidup*, 4(2), 324–338. <https://doi.org/10.38189/jtbh.v4i2.167>
- Septikasari, Z., Retnawati, H., Wilujeng, I., Shaw, R., & Handaka, I. B. (2024). Needs Assessment: Strategy of Integration of Disaster Education at Elementary Schools in Disaster-Prone Areas. *Journal of Education and Learning (EduLearn)*, 18(1), 244–252. <https://doi.org/10.11591/edulearn.v18i1.20921>
- Stemler, S. E., & Naples, A. (2021). *Rasch Measurement v. Item Response Theory: Knowing When to Cross the Line*. <https://doi.org/10.7275/V2GD-4441>
- Struijs, S. Y., Lamers, F., Verdam, M. G. E., van Ballegooijen, W., Spinhoven, P., van der Does, W., & Penninx, B. W. J. H. (2020). Temporal Stability of Symptoms of Affective Disorders, Cognitive Vulnerability, and Personality Over Time. *Journal of Affective Disorders*, 260, 77–83. <https://doi.org/10.1016/j.jad.2019.08.090>
- Subchi, I., Zulkifli, Z., Latifa, R., & Sa'diyah, S. (2022). Religious Moderation in Indonesian Muslims. *Religions*, 13, 451. <https://doi.org/10.3390/rel13050451>
- Sumintono, B., & Widhiarso, W. (2014). *Aplikasi Model Rasch untuk Penelitian Ilmu-Ilmu Sosial (edisi revisi)*.
- Syaifullah, A. (2020). Moderasi Islam dalam Kitab Sabilal Muhtadin: Kearifan Lokal Tanah Banjar. *Mudṣarah: Jurnal Kajian Islam Kontemporer*, 2(1), 31. <https://doi.org/10.18592/msr.v2i1.3676>
- Tayfur, S. N., Prior, S., Roy, A. S., Fitzpatrick, L. I., & Forsyth, K. (2021). Adolescent Psychosocial Factors and Participation in Education and Employment in Young Adulthood: A Systematic Review and Meta-Analyses. *Educational Research Review*, 34, 100404. <https://doi.org/10.1016/j.edurev.2021.100404>
- Trizano-Hermosilla, I., Gálvez-Nieto, J. L., Alvarado, J. M., Saiz, J. L., & Salvo-Garrido, S. (2021). Reliability Estimation in Multidimensional Scales: Comparing the Bias of Six Estimators in Measures with a Bifactor Structure. *Frontiers in Psychology*, 12. <https://doi.org/10.3389/fpsyg.2021.508287>
- Tutz, G. (2024). Invariance of Comparisons: Separation of Item and Person Parameters Beyond Rasch models. *Journal of Mathematical Psychology*, 122, 102876. <https://doi.org/10.1016/j.jmp.2024.102876>
- Uberti, L. J. (2022). Interpreting Logit Models. *The Stata Journal*, 22(1), 60–76. <https://doi.org/10.1177/1536867X221083855>
- Vijayamohan Pillai N & Rjumohan Asalatha. (2020). *Reliability, Validity and Uni-Dimensionality: A Primer*. <https://doi.org/10.13140/RG.2.2.25145.80489>
- Wahyudi, L. E., Mulyana, A., Dhiaz, A., Ghandari, D., Putra Dinata, Z., Fitoriq, M., & Hasyim, M. N. (2022). Mengukur Kualitas Pendidikan di Indonesia. *Ma'arif Journal of Education, Madrasah Innovation and Aswaja Studies*, 1(1), 18–22. <https://doi.org/10.69966/mjemias.v1i1.3>
- Weir, J. P. (2005). Quantifying Test-Retest Reliability Using the Intraclass Correlation Coefficient and the SEM. *The Journal of Strength and Conditioning Research*, 19(1), 231. <https://doi.org/10.1519/15184.1>
- Wyse, A. E. (2010). R.J. DE Ayala (2009). *The Theory and Practice of Item Response Theory*. New York: The Guilford Press. xv + 448 pages. US\$60.00. ISBN: 978-1-59385-869-8. *Psychometrika*,

75(4), 778–779. <https://doi.org/10.1007/s11336-010-9179-z>

Zakariyah, Z., Muhamad Arif, & Nurotul Faidah. (2022). Analisis Model Kurikulum Pendidikan Agama Islam di Abad 21. *At-Ta'dib: Jurnal Ilmiah Prodi Pendidikan Agama Islam*, 1–13. <https://doi.org/10.47498/tadib.v14i1.964>

Zijlmans, E. A. O., Tijmstra, J., Van Der Ark, L. A., & Sijtsma, K. (2019). Item-Score Reliability as a Selection Tool in Test Construction. *Frontiers in Psychology*, 9. <https://doi.org/10.3389/fpsyg.2018.02298>