IJECA (International Journal of Education & Curriculum Application)
Journal homepage: http://journal.ummat.ac.id/index.php/IJECA
ISSN 2614-3380 | Vol. 8, No. 3, December 2025
https://doi.org/10.31764/ijeca.v8i3.35338

The Analysis of Religious Character Instrument using Classical and Modern Theories

Johri Sabaryati[1*], Linda Sekar Utami[1], Siti Ala[2], Pujianti Bejahida Donuata[3], Yulinda Erma Suryani[4]
[1]Physics Education, Muhammadiyah University of Mataram, Indonesia
[2]Discipline of Chemical Engineering, WA School of Mines: Minerals, Energy, and Chemical Engineering, Curtin University, Autralia
[3]Earth Science Education Department, Seoul National University, South Korea
[4]Universitas Negeri Yogyakarta, Indonesia
✉ Author Corresponding: joyafarashy@gmail.com

ABSTRACT

Measuring religious character is essential for understanding how individuals internalize religious values. However, previous studies have mainly focused on conceptual development or instrument construction without systematically comparing psychometric characteristics before and after scaling or integrating Classical Test Theory (CTT) and Item Response Theory (IRT) within a single analytical framework. This study addresses these gaps by examining the characteristics of a multidimensional religious character instrument before and after scaling to ensure score stability and measurement precision. The instrument was a self-report questionnaire using 4–5 point Likert-type items representing five dimensions: intellectuality, ideology, public practice, private practice, and religious experience. A descriptive quantitative design was employed. CTT was used to evaluate item statistics, reliability, and score distribution, while IRT specifically the graded response model (GRM) assessed item functioning across different levels of the latent trait. The summated rating method was applied to transform ordinal responses into standardized scores. Data were collected from 375 students at Widya Dharma University, Klaten, Indonesia, and analyzed using R. The scaling procedure generated positivized z-scores and produced a more compressed score distribution, reflected in decreased mean, standard deviation, mode, and median. Changes in reliability coefficients and the Standard Error of Measurement (SEM) across dimensions indicated that scaling affected measurement precision. GRM analysis confirmed that the instrument effectively discriminated among individuals with low, moderate, and high levels of religious character. Overall, the findings highlight the value of applying scaling procedures and integrating CTT and IRT to improve the accuracy, interpretability, and psychometric robustness of religious character assessments.

Keywords: Religious Character Instrument; Classical; Modern Theories.

Article History:
Received: 01-10-2025
Revised : 30-11-2025
Accepted: 01-12-2025
Online : 13-12-2025

How to Cite (APA style):
Sabaryati, J., Utami, L. S., Ala, S., Donuata, P. B., & Suryani, Y. E. (2025). The Analysis of Religious Character Instrument using Classical and Modern Theories. *IJECA (International Journal of Education and Curriculum Application)*, *8*(3), 344-355. https://doi.org/10.31764/ijeca.v8i3.35338

1. INTRODUCTION

In psychology, constructs that are not directly observable such as religiosity can still be measured through their attributes and behavioral indicators. Contemporary research conceptualizes religiosity as a multidimensional construct encompassing beliefs, attitudes, practices, and lived experiences (Saroglou, 2020). Building on earlier theoretical foundations,

recent studies have reaffirmed that religiosity reflects an interplay of cognitive, ideological, public, private, and experiential components, each contributing uniquely to an individual's overall religious character (Yildiz, 2025). To capture this complexity, measurement instruments frequently employ Likert-type response formats, which remain among the most widely used scaling techniques for assessing psychological attributes.

Religiosity pertains to the quality of an individual's beliefs and attitudes towards the teachings of his religion and religious practices performed within the context of his connection to the creator and fellow beings to attain life's purpose and happinesss (Suryadi, B. & Hayat, 2021). Generally, religiosity has been greatly developed by Western and Eastern experts. Commonly utilized instruments include the Likert and Thurstone scales. The scores obtained from the religiosity character instrument using the Likert scale are presented as ordinal data. Therefore, the measurement results should not be operated arithmetically as is the case using an interval scale (Wu & Leung, 2017). This is due to the limitation of ordinal data, which is suitable for determining mode and median but inadequate for computing the mean and standard deviation for the instrument results. To address the issue, calculate the mean and standard deviation using nonparametric statistic (Setiawati et al., 2013; Wu & Leung, 2017). There is a gap in this research in developing appropriate methods to convert ordinal data into valid interval data that can be processed arithmetically.

Despite their popularity, Likert-type instruments produce ordinal data, meaning that response categories reflect ranked order but do not guarantee equal spacing between points. However, in much of the current psychological and educational measurement literature, researchers often treat ordinal scores as interval-level data, applying arithmetic operations such as computing means, standard deviations, and reliability coefficients without appropriate transformation (Rutkowski, 2025). This practice raises methodological concerns because ordinal data violate key assumptions of parametric analysis. Although several approaches such as score normalization, expanded category scaling, and model-based transformations have been proposed, empirical evaluations of these methods remain underdeveloped. In particular, the summated rating method, which converts raw ordinal scores into standardized z-scores to approximate interval properties, has received limited attention in contemporary psychometric research.

A second methodological issue lies in the limited application of modern measurement theory. While many studies on religious character instruments examine validity or dimensionality, they typically rely solely on Classical Test Theory (CTT). Recent scholarship underscores the importance of integrating Item Response Theory (IRT), especially graded response models, to obtain more precise item-level information and examine how instruments perform across varying levels of the latent trait (Paek & Cole, 2020). Yet, few studies have systematically compared pre- and post-scaling psychometric characteristics using both CTT and IRT. Consequently, there is insufficient understanding of how scaling transformations affect item functioning, reliability, error estimates, and distributional properties particularly within multidimensional religiosity instruments such as those adapted from Huber and Huber's framework. These limitations highlight a clear gap: current assessments of religious character rarely evaluate the consequences of converting ordinal Likert responses into interval scales while simultaneously examining the psychometric implications using both CTT and IRT. This gap is especially pronounced in multidimensional religious character instruments, where each dimension may be differentially affected by scaling procedures.

The assessed attributes will visually represent the assigned scores. To estimate the factors, the scores undergo summation to discern potential differences (León-Mantero et al., 2020). For instance, an instrument of a religious character assessment using a Likert scale with 5 responses, the conversion of values from 1 to 5 implies automatic treatment of responses as numerical, contravening the fundamental assumption of ordinal level measurements (Averin et al., 2017). Therefore, it is necessary to explore methods for converting ordinal data into intervals.

To resolve the issue and approximate the resulting data towards a continuous and normal scale, the Linkert scale points are augmented (Wu & Leung, 2017). The scaling process is a method employed to transform ordinal data into interval data (Setiawati et al., 2013). This occurs because scaling positions the measured object along a a continuum range with a continuous sequence of values (Taherdoost, 2019). In this case, the Likert scale's raw scores are converted into z-scores through a normal distribution, ensuring that the intervals between scores has the same units, as a characteristic of interval data (Setiawati et al., 2013). Traditionally, a Likert scale consists a minimum of three or four points. Expanding it to 11 points (score 0 to 10) allows it to be considered as a continuous measure, permitting the use of arithmetic operations (Awopeju & E. Afolabi, 2016). The instrument utilizes a set of 5 response scales with response choices framed around levels of appropriateness. Therefore, the scaling process is characterized as a response-scaling approach (Setiawati, 2013). Hence, it is essential to assess the scaling results of the Likert-type scale religious character instrument featuring 5 response points using the summated rating method.

Furthermore, the psychometric can be analyzed using classical theory (CTT) and modern theory (Setiawati et al., 2018). This gap highlights the need for further research combining classical and modern theories of measuring religious character, using appropriate methods for converting ordinal data into intervals, and more accurate psychometric analysis to develop instruments that are more comprehensive and can be applied to a wider population. Addressing these issues, the present study analyzes the five-dimensional religious character instrument—covering intellectuality, ideology, public practice, private practice, and religious experience—administered to 375 students at Widya Dharma University, Indonesia. Using the summated rating method, the study transforms ordinal Likert responses into interval-approximated scores and evaluates their psychometric properties through CTT and IRT. The objectives of this research are threefold: (1) to examine how scaling alters score distributions and descriptive statistics; (2) to assess changes in reliability and measurement error before and after scaling; and (3) to evaluate item and person parameters using IRT to determine whether the instrument functions effectively across different levels of religious character. The study makes two key contributions. Theoretically, it advances the discussion on measurement of religiosity by clarifying how scaling procedures influence the statistical and psychometric behavior of multidimensional religious instruments. Methodologically, it demonstrates the value of integrating CTT and IRT in evaluating scaling outcomes, offering a more rigorous framework for transforming ordinal responses into meaningful interval-like data. These contributions highlight the urgency of improving measurement practices to produce more accurate, interpretable, and generalizable assessments of religious character.

2. METHODS

2.1 Research Design

This study employed a quantitative research design to analyze the psychometric characteristics of a multidimensional religious character instrument. The methodological framework followed three structured phases of instrument development preliminary research, prototyping, and product evaluation (Akker et al., 2013), and supported by contemporary instrument development models (Devellis, 2017).

a. Preliminary Research.
   A comprehensive literature review was conducted to identify theoretical foundations of religiosity measurement and evaluate existing multidimensional models. Based on this review, the study adopted the Five-Dimensional Centrality of Religiosity framework developed by Huber & Huber (2012), emphasizing intellectuality, ideology, public practice, private practice, and religious experience.
b. Prototyping.
   In this stage, instrument blueprints and item indicators were constructed by translating each dimension into observable behavioral statements. The initial item pool was reviewed through expert judgment to ensure conceptual alignment, relevance, and clarity. A content validity assessment using the Aiken's V index was conducted with experts in psychology, religious education, and measurement.
c. Product Evaluation.
   The validated prototype was administered in a large-scale trial to examine construct validity, reliability, and item functioning. The study employed both Classical Test Theory (CTT) and Item Response Theory (IRT) to obtain comprehensive psychometric information. Following data collection, a summated rating scaling procedure was applied to transform ordinal Likert scores into interval-approximated values, allowing for more precise statistical analysis.

2.2 Respondents

The research was conducted at Widya Dharma University Klaten, North Klaten District, Klaten Regency, Central Java Province. The target population consisted of undergraduate students across 17 academic programs. A proportionate stratified sampling technique was used to ensure representation from each program, aligning with recommendations for heterogeneous educational populations (Creswell, 2012). The subjects of this research were the students of Widya Dharma University Klaten majoring in Indonesian Language and Literature Education, English Education, Regional Language and Literature Education, Geography Education, Pancasila and Citizenship Education, Mathematics Education, Elementary School Teacher Education, Management, Accounting, Tax Management, Civil Engineering, Electrical Engineering, Agricultural Product Technology, Psychology, Physiotherapy, Informatics Engineering, and Informatics Management. There were 375 student respondents in total, with male students comprising 31%, which equals to 116 students. The remaining students were female.

2.3 Instruments

The item creation process begins by gathering various theories related to the definition of religion. They were analyzed based on the field-specific conditions. An instrument suitable for students by adopting Huber & Huber's (2012) approach. Religiosity is observed through the dimensions of intellectual (religious knowledge), ideology (belief), public practice, private

practice, and religious experience. Students are presented with five response options to assess the character of religiosity in response to a statement: STS = *Sangat Tidak Sesuai* (Very Inappropriate), TS = *Tidak Sesuai* (Inappropriate), CS = *Cukup sesuai* (Quite Appropriate) S = *Sesuai* (Appropriate) and SS= *Sangat Sesuai* (Very Appropriate). The survey opted for a Likert scale due to its suitability for measuring behavior such as religious characters (Price, 2017). This deep data analysis study uses statistical calculations with the help of Microsoft Excel, SPSS 26, and R programs. The calculations compare the results of KMO MSA, Eigen Value, reliability, and Standard Error of Measurement (SEM) between raw scores (original) and standardized scores (rescaling) through summated rating scaling.

3. RESULT AND DISCUSSION

3.1 Psychometric Characteristics of Pre-Scaling and Post-Scaling Data

The Likert type religious character instrument is scaled using the summated rating method. The data which constitutes student responses is processed using the Microsoft Excel program. The process involves several stages: (1) the calculation begins by determining the frequency (f) for each response scale per statement item, (2) dividing the frequency by the number of respondents (n) to obtain the proportion (p) for each item, (3) establishing cumulative proportions (pk) by summing the previous proportions, (4) determining the middle pk (pk-t) using the formula ½p + pkb or half the proportion in that category added with the cumulative proportion of the previous category, (5) determining the deviation value (z) by converting the pk-t score into a z score by referring to the normal curve z table,(6) determining the smallest deviation by cumulatively adding scores until reaching the minimum value of 0, and (7) rounding the results from point 5 (Setiawati et al., 2018). The pre-scaling data are displayed in the following matrix. Table 1 through 5 display examples of calculations for the 5 items (specifically items 1).

Table 1. Summated Rating Scaling Calculation for Item B1

Response Category	f	p	pk–t	Z	Adjusted Z	Final Score
1	1. 1	2. 0.003	0.0015	–3.00	0.00	0
2	0	0.000	0.0030	–2.75	0.25	0
3	38	0.101	0.0535	–1.60	1.40	1
4	193	0.515	0.3615	–0.35	2.65	3
5	143	0.381	0.8095	0.88	3.88	4

Sources: Personal data (2025)

By considering the process of scaling the instrument through the methods described in Table 1, it can be concluded that the impact of response score variation on the summated rating scale includes finer data granularity, changes in descriptive statistics (such as mean, variance, and data normality), validity and reliability results, and interpretation of research results. In addition, scaling Likert-type instruments with the summated rating method is actually a scaling process with a response approach. In this study, an attempt was made to compare the results of KMO MSA, Eigen Value, reliability, and Standard Error Measurement (SEM) raw scores (original) and standardized scores (rescaling). The instrument was scaled. In classical theory, instruments are analyzed using the summated rating method.

The reliability of the instrument in this study was calculated using the Cronbach's Alpha formula. Reliability is a psychometric trait frequently applied in classical theoretical approaches.

The initial pre-scaling data's instrument reliability estimation resulted in an alpha coefficient of 0.87, which increased to 0.92 at post-scaling. The reliability of the instrument is achieved when the alpha coefficient value is ≥ 0.65 (Hanson, 1973). The reliability estimation results for the religious character instrument for pre-scaling and post-scaling were more than 0.65. Therefore, it can be concluded that the religious character instrument developed was reliable. The reliability coefficient for pre-scaling and post-scaling of the religious character instrument with 5 scale experienced a change, although it was not significant. The reliability coefficient tends to increase for post-scaling with summed rating due to the scores have been standardized or scaled. The reliability estimation are presented in Table 2 below.

Table 2. Results of the Reliability Estimation for the Religious Character Instrument

	SD	Alpha	SEM
Initial Data	8.800791	0.871992	3.148768
SRS	9.37052	0.915802	2.719034

Sources: Personal data (2025)

Table 2 also illustrates a decrease in the Standard Error of Measurement (SEM) value for the pre-scaling and post-scaling data. SEM on the post-scaling data is lower compared to SEM on the pre-scaling data. Meanwhile, the standard deviation value experienced an increase, although it was not significant.

3.2 Scaling Using IRT Approach

Scaling with a modern approach is related to the employed measurement model. In this research, a graded response model (GRM) is employed. Within this model, two parameters additionally impact the score transformation during the scaling process: the differential power parameter and the item difficulty index, commonly referred to as the *index probability of endorsemet* (Embretson S. E., 2000) in non-cognitive assessments, as shown in Table 3.

Table 3. Comparison of Pre-scaling and Post-scaling GRM Fit (p.S_X^2)

Item	Pre-Scaling	Post-Scaling	Interpretation
B1	0.58	0.68	Fit improves
B2	0.40	0.53	Fit improves
B3	0.41	0.56	Fit improves
B4	0.43	0.33	Slight decrease, still acceptable
B5	0.28	0.53	Fit improves
B6	0.03	0.01	Misfit persists
B7	0.00	0.00	Misfit persists
B8	0.00	0.02	Still misfit
B9	0.39	0.48	Fit improves
B10	0.26	0.15	Slight decrease
B11	0.22	0.07	Slight decrease
B12	0.01	0.04	Still misfit
B13	0.22	0.26	Stable/slight improvement
B14	0.69	0.56	Both acceptable
B15	0.30	0.44	Fit improves
B16	0.26	0.47	Fit improves
B17	0.00	0.01	Still misfit

Item	Pre-Scaling	Post-Scaling	Interpretation
B18	0.09	0.39	Fit improves
B19	0.73	0.43	Slight decrease, still acceptable
B20	0.11	0.35	Fit improves
B21	0.45	0.50	Stable/slight improvement
B22	0.00	0.00	Misfit persists
B23	0.45	0.14	Decrease
B24	0.00	0.01	Misfit persists
B25	0.12	0.17	Slight improvement
B26	0.28	0.38	Fit improves
B27	0.50	0.64	Fit improves
B28	0.28	0.19	Slight decrease
B29	0.06	0.03	Still borderline
B30	0.35	0.44	Fit improves

The comparison of item fit statistics based on the Graded Response Model (GRM) before and after scaling demonstrates several important changes in item functioning. Overall, the majority of items show improved model fit following the application of the summated rating scaling procedure. Before scaling, several items (B6, B7, B8, B12, B17, B22, B24, and B29) exhibited poor fit, as indicated by p.S_X^2 values below 0.05. These items remained problematic after scaling, suggesting that misfit may be attributable to factors intrinsic to item wording, content alignment, or response category functioning rather than to the scaling transformation itself. Despite this, most items experienced a clear improvement in fit after scaling. For instance, items B1–B5, B9, B13–B16, B18, B20–B21, B25–B27, and B30 showed increases in p.S_X^2 values, indicating a better alignment between the observed responses and the GRM model assumptions. A smaller number of items experienced minor decreases in fit after scaling (such as B4, B10, B11, B19, B23, and B28), but these decreases remained within acceptable thresholds (p.S_$X^2 \geq 0.05$). This indicates that the scaling process did not substantially diminish item performance for these indicators. Items with already high pre-scaling fit, such as B14, B19, and B27, maintained stable or acceptable fit after scaling. Overall, the results suggest that the summated rating method enhances the psychometric functioning of most items in the religious character instrument by improving the distributional characteristics of the data and strengthening the alignment with the GRM model. However, several items consistently demonstrating misfit should be reviewed further for potential revision, rewording, or removal to improve the overall measurement quality. Table No.3 displays the results of fit data analysis using the GRM, comprising 23 fit data from pre-scaling and 22 fit data from post-scaling (Taherdoost, 2019). An item is considered fit if the p-value of p.S_X2 is greater than 0,05. Subsequently,an analysis of the fit item parameters was conducted. The analysis results can be represented by a curve for each item, illustrated in the following figures. The following Figure 1 and Figure 2 illustrate the results of unscaled data.

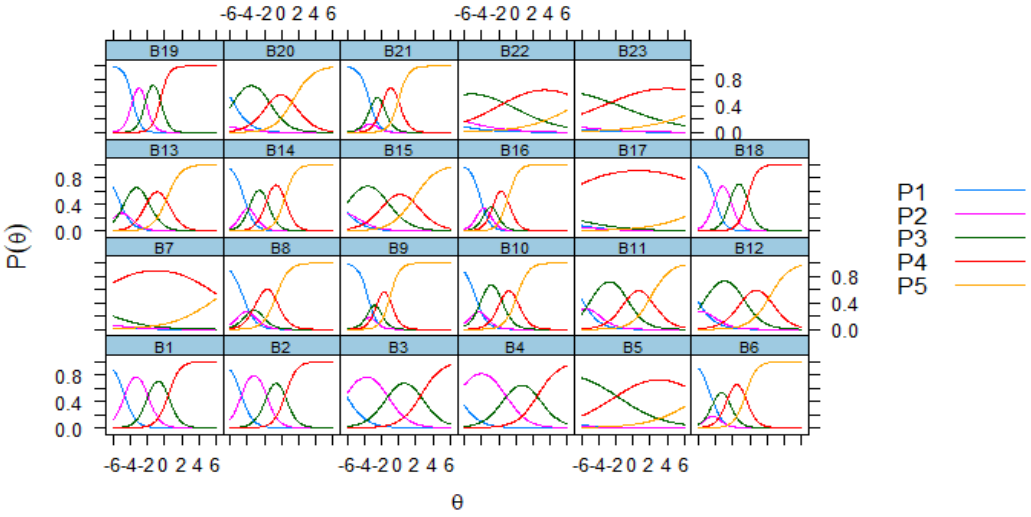

Figure 1. Item characteristic curve (ICC) for pre-scaling data

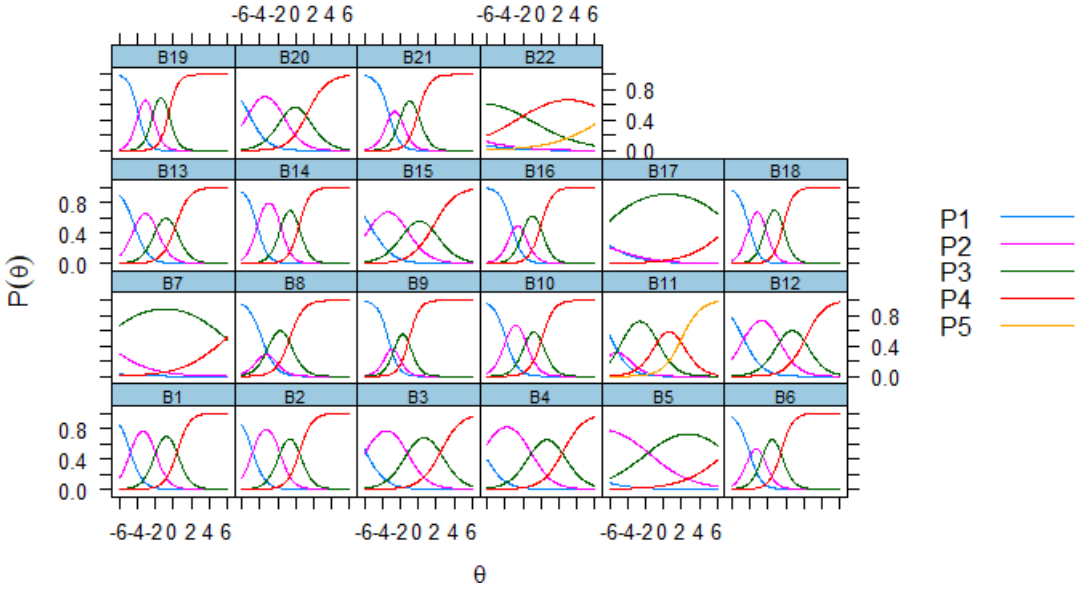

Figure 2. Item characteristic curve (ICC) for post-scaling data

The subsequent analysis concerns the instrument profile, specifically focusing on examining the information function's value and standard error. This analysis aimed to assess the compatibility of the instrument with the abilities of the respondents, demonstrated in Figure 3 below.

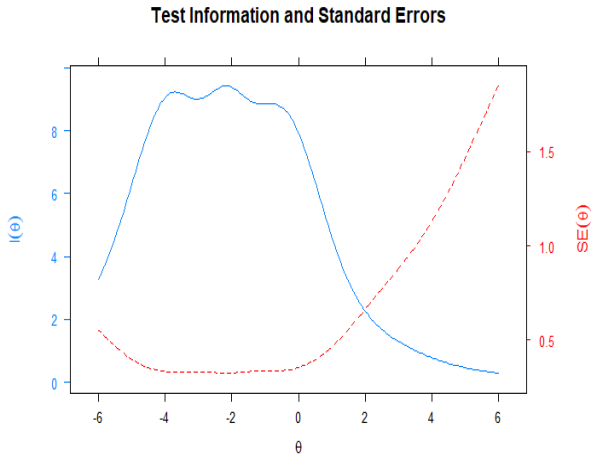

Figure 3. Information function curve and standard error pre-scaling

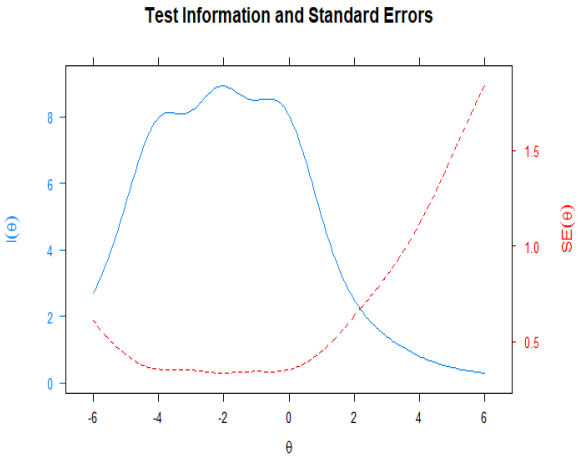

Figure 4. Information function curve and standard error post-scaling

Scaling using the summed rating method resulted in a clearer curve for the items, which better reflects the structure of the GRM. In the post-scaling diagram, most items showed more regular response categories with distinct category peaks and a more consistent threshold sequence compared to the pre-scaling representation. Extreme categories became more stable and no longer dominated the entire ability range, indicating a more balanced response distribution. However, some items (e.g., B7, B17, B22) continued to exhibit suboptimal curves both before and after scaling. This suggests that the problem lies in the quality of the items themselves, rather than in the scaling process. Overall, scaling had a positive impact on the functionality of the items in the GRM and improved the instrument's ability to more accurately differentiate the degree of religiosity among students.

In this study, the religious character instrument employed a Likert-type scale featuring 5 response points. This type of scale has long been used and comprises 5 equally distributed and balanced points (Taherdoost, 2019). Each statement within the religious character instrument utilized in this study offers 5 response choices: 1 for "Very Inappropriate," 2 for "Not Appropriate," 3 for "Quite Appropriate," 4 for "Appropriate," and 5 for "Very Appropriate." Assessing religious character through a Likert-type instrument results in an ordinal score. The research findings on religious character instruments showed a decrease in SEM post-scaling. To gain precision and accuracy involves aggregating the items into a large total score to minimize measurement errors. To minimize the measurement error, Likert scale data is transformed into interval data with suitable weighting (Barge, 1988). One way to achieve this is converting ordinal data into interval data (Setiawati, 2013). To convert this data, one method involves employing the summated rating technique on the religious character instrument, namely (1) determining the frequency (f) of each response scale for each statement item, (2) dividing the frequency by the number of respondents (n) to obtain the proportion (p) for each item, (3) establishing cumulative proportions (pk) by summing the previous proportions, (4) determining the middle pk (pk-t) using the formula ½p + pkb or half the proportion in that category added with the cumulative proportion of the previous category, (5) determining the deviation value (z) by converting the pk-t score into a z score by referring to the normal curve z table,(6) determining the smallest deviation by cumulatively adding scores until reaching the minimum value of 0, and (7) rounding the results from point 5 (Azwar, 2016).

The result of this summated rating is a new score, which is z score after rounding. The conducted data analysis produces a change in the score for each response in the z rounding score, indicating that the response score for each item indicating differences from those of pre-scaling (Setiawati, 2013). It is evident that grouping the response to each item in rounding the z-score is challenging due to the varying results shown for each item. Examining the increase in the standard deviation shows that the average data is getting higher due to how students respond rhythmically to this item, it is noticeable that the answers predominantly lean towards a positive response. This religiosity character instrument has a high reliability value of ≥ 0.65 for pre-scaling and post-scaling (Hanson, 1973). This indicates that the religiosity character instrument is reliable, ensuring consistent data output within the same sample over time (Wadkar et al., 2016). In addition, the reliability of SEM data post-scaling decreased by 2.719034 from pre-scaling, which was 3.148768. The size of the SEM is influenced by the variance. The higher the variance value, the higher the SEM value. Furthermore, an increased number of respondents will lead to greater variance, consequently resulting in a higher SEM value as well. A higher SEM indicates increased variability within the data (Frey, 2018). The analysis indicates that among the items examined, the p.S_X2 value is greater than 0.05, which means there is a higher proportion of items that do not fit after scaling. The item parameter values are structured with b1 < b2 < b3, allowing for the measurement of item difficulty across different ability ranges for participants. Both pre-scaling and post-scaling data are derived from the information function curve points.

4. CONCLUSION

In this section the author details the conclusions of the results of the discussion and data analysis and is advised to submit further research to the next researcher. Based on the findings, this study concludes that the application of the summated rating method contributes meaningfully to enhancing the quality of measurement outcomes in the assessment of religious character. By converting ordinal Likert responses into positivized z-scores, the method facilitates a more standardized and interpretable scoring framework across dimensions of religiosity. This transformation helps reduce distortions commonly associated with treating ordinal responses as interval-level data, thereby improving comparability of scores and supporting more accurate descriptive and inferential analyses.

Methodologically, the study provides empirical evidence on how scaling affects psychometric indicators, particularly reliability coefficients and the Standard Error of Measurement (SEM). The observed shifts between pre- and post-scaling phases demonstrate that scaling procedures can influence both the internal consistency and precision of an instrument. These findings highlight the importance of reconsidering standard practices within psychological and educational measurement, where ordinal data are often analyzed using interval-based statistics without transformation. Moreover, by integrating Classical Test Theory (CTT) and Item Response Theory (IRT), this study strengthens the theoretical bridge between traditional and modern measurement perspectives. The GRM results indicate that the instrument is capable of distinguishing respondents across varying levels of religiosity, reinforcing its potential utility for diverse populations and educational contexts. However, these contributions should be interpreted within the study's methodological boundaries. First, the instrument was tested within a single university context, which may limit the generalizability of the findings across cultural or demographic groups. Second, although the summated rating method improved several psychometric indicators, some items continued to display misfit within the GRM framework,

suggesting that scaling alone cannot correct item-level shortcomings related to content, clarity, or construct representation. Third, the study relied on self-report measures, which remain vulnerable to social desirability bias in religious contexts.

Future research should therefore aim to replicate the analysis in broader and more diverse samples, incorporate qualitative validation to refine problematic items, and compare multiple scaling approaches such as Rasch-based transformations or ordinal logistic modelling to evaluate their relative effectiveness. Researchers may also extend this work by examining longitudinal stability of scaled scores or by integrating multimethod assessments (e.g., behavioral indicators, peer evaluations) to reduce bias inherent in self-reported religiosity. In summary, this study advances both measurement theory and the practice of religiosity assessment by demonstrating the value of scaling techniques and the integration of CTT and IRT in strengthening psychometric robustness. The findings underscore the need for more rigorous methodological attention when converting ordinal responses into interpretable interval-like measures, while offering a practical framework for improving the validity, reliability, and fairness of religiosity measurement tools.

ACKNOWLEDGEMENT

The authors would like to express their deepest gratitude to all individuals and institutions who contributed to the completion of this research. Special thanks are extended to the academic mentors and colleagues who provided invaluable guidance and constructive feedback throughout the research process. We also thank the educational institutions and participants involved in the data collection for their cooperation and willingness to share insights. Finally, we are grateful for the support of our families and friends, whose encouragement motivated us to persevere in completing this study.

REFERENCES

Akker, J. Van den, Bannan, B., Kelly, A. E., Nieveen, N., & Plomp, T. (2013). *Educational Design Research*.http://www.eric.ed.gov/ERICWebPortal/recordDetail?accno=EJ815766

Averin, A. D., Yakushev, A. A., Maloshitskaya, O. A., Surby, S. A., Koifman, O. I., & Beletskaya, I. P. (2017). Synthesis of porphyrin-diazacrown ether and porphyrin-cryptand conjugates for fluorescence detection of copper(II) ions. *Russian Chemical Bulletin*, *66*(8), 1456–1466. https://doi.org/10.1007/s11172-017-1908-3

Azwar, S. (2016). *Dasar-Dasar Psikometrika (Edisi II).* Pustaka Pelajar.Yogyakarta.

Barge, M. (1988). A method for constructing attractors. *Ergodic Theory and Dynamical Systems*, *8*(3), 331–349. https://doi.org/10.1017/S0143385700004491

Csikszentmihalyi, M. (2014). *Flow and the Foundations of Positive Psychology* (pp. 1–298). Springer.https://doi.org/10.1007/978-94-017-9088-8

Creswell, J. W. (2012). *Educational Research(Planning, conducting and evaluating quantitative and qualitative Research)* (Vol. 148). Pearson Education, Inc.

Devellis, R. F. (2017). *Scale Development Theory and Applications*. SAGE Publications, Inc.

Embretson S. E., &Reise S. P. (2000). *Item response theory for psychology.* Elbaum Associates, Publisher.

Frey, B. B. (2018). Variance. *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*, *August*. https://doi.org/10.4135/9781506326139.n737

Gravemeijer, K., & Cobb, P. (2013). *Educational Design Research*. Enschede: SLO Netherlands Institute for Curriculum Development. https://www.slo.nl

Hanson, R. A. (1973). Essentials of educational measurement. *Journal of School Psychology*, *11*(2), 172–173. https://doi.org/10.1016/0022-4405(73)90057-5

Kampen, J. K. (2019). Reflections on and test of metrological properties of summated rating, Likert, and other ordinal scales. *Measurement*, 137, 428–434.

https://doi.org/10.1016/j.measurement.2019.01.083

León-Mantero, C., Casas-Rosal, J. C., Pedrosa-Jesús, C., & Maz-Machado, A. (2020). Measuring attitude towards mathematics using Likert scale surveys: The weighted average. *PLoS ONE*, *15*(10 October), 1–15. https://doi.org/10.1371/journal.pone.0239626

Awopeju, O. A., & Afolabi, E. R. I. (2016). Comparative analysis of classical test theory and item response theory based item parameter estimates of senior school certificate mathematics examination. *European Scientific Journal*, *12*(28), 263-284. https://doi.org/10.19044/esj.2016.v12n28p263

Paek, I., & Cole, K. (2020). *Using R for Item Response*. Routledge.

Price, L. R. (2017). *Psychometric Methods Methodology in the Social Sciences*. The Guilford Press. www.guilford.com/MSS

Rutkowski, I. P. (2025). Steven's Measurements Scales In Marketing Research – A Continuation Of Discussion On Whether Researchers Can Ignore The LIKERT Scale's Limitations. *Sciendo*, *55*(1), 39–55. https://doi.org/10.2478/minib-2025-0003

Saroglou, V. (2020). *The Psychology of Religion (1st ed.).* Routledge. https://doi.org/https://doi.org/10.4324/9781351255967

Setiawati, F. A., Izzaty, R. E., & Hidayat, V. (2018). Analisis Respons Butir Pada Tes Bakat Skolastik. *Jurnal Psikologi*, *17*(1), 1. https://doi.org/10.14710/jp.17.1.1-17

Setiawati, F. A., Mardapi, D., & Azwar, S. (2013). Penskalaan Teori Klasik Instrumen Multiple Intelligences Tipe Thurstone Dan Likert. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *17*(2), 259–274. https://doi.org/10.21831/pep.v17i2.1699

Suryadi, B. & Hayat, B. (2021). *Religiusitas: Konsep, Pengukuran, dan Implementasi di Indonesia*. Bibliosmia Karya Indonesia.

Taherdoost, H. (2019). *What Is the Best Response Scale for Survey and Questionnaire Design; Review of Different Lengths of Rating Scale / Attitude Scale / Likert Scale by Hamed Taherdoost :: SSRN*. *8*(1), 1–10. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3588604

Wadkar, S. K., Singh, K., Chakravarty, R., & Argade, S. D. (2016). Assessing the Reliability of Attitude Scale by Cronbach's Alpha. *Journal of Global Communication*, *9*(2), 113. https://doi.org/10.5958/0976-2442.2016.00019.7

Wu, H., & Leung, S. O. (2017). Can Likert Scales be Treated as Interval Scales?—A Simulation Study. *Journal of Social Service Research*, *43*(4), 527–532. https://doi.org/10.1080/01488376.2017.1329775

Yildiz, K. (2025). Exploring the Theoretical Connections Between Psychology of Religion and Religious Pedagogy : An In-depth Analysis. *Turkish Journal for the Psychology of Religion*, *11*(June), 145–178. https://doi.org/https://doi.org/10.4324/9781351255967