JURNAL
PENELITIAN
Volume 20 Nomor 1 2023

ISSN 1829-9903 (Print)
2541-6944 (Online)

Submitted: 17/02/2023
Reviewed: 02/03/2023
Approved: 27/03/2023

Url Website: http://e-journal.uingusdur.ac.id/index.php/Penelitian/
Url DOI: https://doi.org/10.28918/jupe.v20i1.1098

Psychometric Characteristics of the *Kalam* Cognitive Test (KCT): An Andrich Rating Scale Analysis

Kriswantoro
Sekolah Tinggi Agama Islam Ma'arif Jambi
kriswantoro.staima@gmail.com

Rizki Nor Amelia
Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Semarang
rizkinoramelia@mail.unnes.ac.id

Abstract:
The science of Kalam is one of the compulsory courses for prospective Islamic teachers, and mastery of this material is very important. Therefore, the measurement instrument used must have good psychometric characteristics to accurately reflect the abilities of prospective Islamic teachers. The purpose of this study is to analyze the psychometric characteristics of the Kalam Cognitive Test (KCT), which has been developed with the Andrich Rating Scale and assisted by the Winsteps program. The research, which was conducted in the spring semester of the 2022–2023 academic year, involved 44 prospective Islamic religion teacher students in the Islamic Religious Education Study Program at Jambi Ma'arif Islamic College who were selected through cluster random sampling techniques. The results of the analysis show that the KCT has good psychometric characteristics in terms of fulfilling the assumptions of the analysis, item reliability, person reliability, and item difficulty level. Related to the rating scale analysis, it can also be concluded that all rating scale categories have met the required criteria so that they have functioned properly.

Keywords: *psychometric characteristics, kalam cognitive test, andrich rating scale*

INTRODUCTION

The science of *Kalam* is an object of study in the form of science in Islam that is studied using the basis of thinking in the form of logic and the basis of personal or group beliefs to answer questions about the existence of God, how God is, what his form is like, and other similar questions related to God. The main purpose of the science of *Kalam* is to explain the basis of the Muslim faith in a philosophical and logical order. For believers, the evidence of God's existence and everything related to God that is in the Qur'an, Hadith, the words of companions who heard the Prophet's words directly, and so on, is sufficient. But when this problem is faced in a wider and more open world, then the naqli arguments do not play a role. This is because not everyone believes in the truth of the Qur'an. Therefore, it is necessary to interpret the reasoning of the dalil that already exists in the Qur'an to explain it. Initially, the conversation about theology was just an ordinary debate to sharpen Islamic understanding, but over time it formed a pro-contra group that fought hatred, hostility, and even war.

The science of *Kalam* talks about the existence of Allah, the attributes that must exist in Him, the attributes that do not exist in Him, and the attributes that may exist in Him, and talks about the Messengers of Allah to establish the truth of their apostleship and know the attributes that must exist in them, the attributes that cannot exist in them, and the attributes that may exist in them (Hasbi, 2015). The science of *Kalam* is one of the compulsory courses in the curriculum structure of the Jambi Ma'arif Islamic College that must be taken by prospective Islamic Religious Education teachers. This course aims to equip mastery of theory and practice and strengthen faith or belief in Allah SWT through reason regarding material and characteristics: Introduction to The science of *Kalam* (Islamic Theology); Basics and History of the Emergence of The science of *Kalam* Relationship between The science of *Kalam*, Philosophy, and Sufism; Faith, Kufr, Nifaq, and Shirk; Tawheed; Khawarij and Murji'ah; Jabariyah and Qadariyah; Theological Thought of al-Mu'tazilah and al-Syi'ah; Ahlussunnah Salaf and Khalaf; Comparison of thoughts in The science of *Kalam*; and the Pillars of Faith One of the ways that can be used to measure mastery of the The science of *Kalam* course is through an essay test.

Essay tests are generally divided into two types: restricted response items and extended response items. Restricted response items are used to measure students' ability to explain cause-and-effect relationships, explain the application of a concept, present relevant arguments, formulate hypotheses, make assumptions about conclusions, explain data limitations, and explain methods and procedures; while extended response items are used to measure student's ability to produce, organize, and express ideas to integrate learning experiences in different contexts (Baek, 2003). Therefore, essay items are a unique test format because they can assess students' ability to generate and construct answers, communicate ideas in writing, and require students to respond rather than choosing one of the alternatives provided (Tozoglu, Tozoglu, Gurses, & Dogar, 2004). Although essay tests are easy to construct and relatively valid for use as tests of high cognitive processes, the reliability of this type of test is quite difficult to estimate (Tuckman, 1993). Another disadvantage of essay questions is the possibility of high subjectivity and many factors that may be unrelated to the answer, thus impacting the test score (Gupta, Jain, & D'souza, 2016).

One way that can be used to improve the reliability of essay tests and reduce the subjectivity factor is through the use of scoring guidelines (rubric) (Wahyuni, Gumela, & Maulana, 2021), which contain a point scale with specific descriptors used to evaluate student answer responses (Sumekto & Setyawati, 2018), where the points generally range from zero to four (a 4-point rating scale) (Martinez, 1997). The rating scale model is a latent structure model for a polytomous response to a set of items that have the same category score (Andersen, 1997; Clark et al., 2020). Classical Test Theory is historically the oldest test theory for rating scales, but in many cases, it has been replaced by Modern Test Theory (i.e. Item Response Theory, Rasch Model) (Andrich, 2011) due to limitations in (1) assuming a linear relationship between latent variables and observed scores, which rarely represents the empirical reality of behavioral constructs; (2) true scores cannot be estimated directly or only by making assumptions that are difficult to meet; and (3) parameters such as reliability, discrimination, location, or factor loadings depend on the sample used (Rusch, Lowry, Mair, & Treiblmaier, 2007).

The Andrich Rating Scale is one of the models in Modern Test Theory that is suitable for granulated items that are multi-graded (Cordier et al., 2018) and can be used to determine the rating scale, both in terms of the number of options and their labels (Zile-Tamsen, 2017). If an item fits this model, the item parameters, represented in its difficulty, can be assessed independently of individual ability, and individual ability can also be assessed independently of item difficulty (Lang & Tay, 2021). An important characteristic of the Andrich Rating Scale is that it provides a means of estimating the measurement stability of items within item thresholds and that a single parameter is established across a sample of individuals (i.e., parameter stability). In the current context, the consistent assessment provides a way to determine the extent to which item ratings or individual item parameters show changes in rating events (Youssef, 2022). In Indonesia, research that explores the psychometric characteristics of the science of *Kalam* cognitive test using the Andrich Rating Scale is quite difficult to find. Therefore, this study was conducted to describe the psychometric characteristics of *Kalam* Cognitive Test (KCT) instrument using the Andrich Rating Scale.

Therefore, this research explores a crucial aspect of education by focusing on the assessment tool used in the field of *Kalam*, a compulsory subject for prospective Islamic teachers. The novelty of this research lies in its examination of the psychometric characteristics of the *Kalam* Cognitive Test (KCT), which is essential for assessing the competence of future Islamic educators. This research addresses a critical gap in the literature by specifically investigating the assessment tools used in the domain of Islamic education, which is an under-explored area in the field of psychometrics. This delves into the unique attributes of the KCT, emphasizing its importance as a specialized cognitive assessment tool tailored to measure the competence of prospective Islamic teachers in *Kalam*. The article's novelty lies in its rigorous psychometric analysis, utilizing the Andrich Rating Scale model, to evaluate the effectiveness and reliability of the KCT. This analytical approach is less commonly employed in the study of educational assessment tools, making it a distinctive feature of this research.

This research also recognizes the broader educational implications of the KCT's psychometric characteristics. It highlights how the quality and accuracy of the assessment tool can significantly impact the preparation and competence of future Islamic teachers, ultimately affecting the quality of Islamic education. This not only addresses a novel and specific need in the domain of Islamic education but also positions itself as a pioneering study in the broader context of assessment tool development and psychometric analysis. Its findings have the potential to impact both educational practice and research across various disciplines.

In general, this research is descriptive quantitative and is conducted in the odd semester of the 2022/2023 academic year. The population in this study were 71 prospective Islamic Religious Teachers in Semester VII of the Islamic Religious Education Study Program at Ma'arif Jambi Islamic College who received The Science of *Kalam* and Its Development course, which was divided into 3 study groups. From this population, the cluster random sampling technique was used so that a sample of two study groups was selected (N = 44, female = 26; male = 18). The instrument developed is the science of *Kalam* Cognitive Test Instrument (see Table 1) in the form of a limited description question of 10 items with a minimum score of 0 and a maximum score of 5. In scoring, a rubric is used that aims to (1) improve scoring consistency, (2) serve as an assessment guide for different assessors, and (3) provide a measure of validity for complex ideas (Jonsson & Svingby, 2007).

Table 1. *Kalam* Cognitive Test Specification Table

Subject Matter	Indicator	Cognitive Dimension	Item Number
Introduction to the science of *Kalam* (Islamic Theology)	Presenting the concept and role of *Kalam* science in social life	C3	1
Basics and history of the emergence of *Kalam* science	Analyzing the history of the development of *Kalam* Science during the time of Khalifah al-Rasyidin, Bani Umayyah, and Bani Abbasiyah	C4	2
The relationship between *Kalam*, philosophy, and sufism	Discovering the relationship between *Kalam*, philosophy, and sufism	C4	3
Iman, Kufr, Nifaq, Syirk	Examining the levels of syirk	C4	4
Tauhid	Explaining the meaning of Tauhid Rububiyyah and Tauhid Uluhiyah	C3	5
Jabariyah and Qadariyah	Comparing the main thought of Jabariyah and Qadariyah	C5	6
Khawarij and Murji'ah	Interpreting the propositions that form the basis of the Murji'ah	C5	7
Theological thought of al-Mu'tazilah and al-Syi'ah	Analyzing the main thought of al-Muktazilah and al-Syi'ah	C4	8
Ahlussunnah Salaf and Khalaf	Analyzing the arguments of the Koran which became the basis of the Salaf dan Khalaf	C4	9
Main issues discussed in the Science of *Kalam* (Comparison of thoughts)	Summarizing the views held by the moderate and extreme groups based on data	C5	10

In this context, these levels, often referred to as Bloom's Taxonomy, represent a hierarchy of cognitive skills that help educators and learners understand the depth of thinking required for different tasks and objectives, where C1 represents remembering level, C2 is understanding level, C3 is applying level, C4 is analyzing level, C5 is evaluating level, and C6 is creating level.

Data in the form of polytomous answer responses were processed with the help of the Winsteps Rasch Model program with the Andrich Rating Scale formulation presented in equation (1).

$$\Pr\{X_{ni=}\mathrm{x}\} = \frac{\exp\left(z_{ix}(\beta_n)\right)}{\sum_{k=o}^{m} exp\left(z_{ix}(\beta_n)\right)} \quad \dots (1)$$

where $z_{ix}(\beta_n) = a_{ix} + b_{ix}\beta_n$ and $a_{ix} = -\sum_{k=0}^{x} \delta_{ix}; b_{ix} = x_i; \delta_{i0} \equiv 0$

In this equation, The Rasch Model expressed in terms of thresholds δ_{ix} and b_{ix} = x is an integer scoring function for successive categories (Andrich, 2011)

DISCUSSION

The Andrich Rating Scale analysis starts by proving the assumptions on the KCT. In the Rasch Model, two assumptions are of primary concern, namely unidimensionality (whether the items form one common latent trait) and local independence (the likelihood of a person answering correctly on an item independent of other items) (Ajeigbe & Afolabi, 2017). Fit statistics, showing the degree to which the pattern of observed answers and the modeled expectations are examined in terms of item fit and person fit in the Rasch model, are used to determine unidimensionality and local independence (Sick, 2010). Generally, a variance of more than 20% has been approved to declare the unidimensionality of an instrument (Bakar, Maat, & Rosli, 2019), but for Rasch Model analysis, the minimum criteria for raw variance explained by measures are 40% and unexplained variance in the first contrast is no more than 15% (Aziz, Masodi, & Zaharim, 2013). Based on these criteria, Table 2 shows that the assumption of unidimensionality has been met because the raw variance explained by measures was 44.4%, which was 0.7% less than the modeled percentage. The eigenvalue of unexplained variance in the first contrast shows 2.1 units and represents 11.9%, which can be accepted as being less than 15%. According to DeMars (2010), the assumption of local independence will be automatically fulfilled when the assumption of unidimensionality is also fulfilled. From this, it can be concluded that both assumptions required to build the construct validity of the Rasch model have been met, so that the analysis can continue.

Table 2. Standardized Residual Variance (in Eigenvalue units)

		Empirical		**Modeled**
Total raw variance in observations	=18.0	100.0%		100.0%
Raw variance explained by measure	=8.0	44.4%		45.1%
Raw variance explained by persons	=3.5	19.5%		19.8%
Raw variance explained by items	=4.5	25.0%		25.3%
Raw unexplained variance (total)	=10.0	55.5%	100.0%	54.9%
Unexplained variance in 1st contrast	=2.1	11.9%	21.4%	
Unexplained variance in 2nd contrast	=1.8	10.0%	18.1%	
Unexplained variance in 3rd contrast	=1.6	8.8%	15.8%	
Unexplained variance in 4th contrast	=1.3	6.9%	12.5%	
Unexplained variance in 5th contrast	=0.9	5.0%	9.0%	

After both assumptions are met, the next step is to estimate reliability. Reliability is the consistency or stability of measurement. A test or instrument with good reliability means that respondents will obtain the same score on retesting as long as no other extraneous factors affect the score (Segal & Coolidge, 2018). In the Rasch Model, person reliability is equivalent to traditional test reliability where low values indicate a narrow range of person measures, or a small number of items; whereas item reliability has no traditional equivalent where low item reliability means that the sample size is too small for stable item estimates based on the current data.

The summarized analysis results, as presented in Table 3, unequivocally affirm the *Kalam* Cognitive Test (KCT) instrument's commendable reliability coefficient. This finding underscores the consistency and dependability of the instrument in measuring the cognitive abilities related to *Kalam*. A strong reliability coefficient indicates that the KCT consistently produces similar results when administered to the same group of individuals, thereby enhancing its trustworthiness as a tool for assessing the competence of prospective Islamic teachers in this subject. This reliability attribute is pivotal in educational assessment, as it ensures that the KCT yields dependable and replicable outcomes, further enhancing its utility and credibility within the realm of Islamic education.

Table 3. Item and Person Reliability

Real RMSE		**Model RMSE**	
Item Reliability	Separation	Item Reliability	Separation
0.92	3.28	0.92	3.42
Person Reliability	Separation	Person Reliability	Separation
0.80	2.01	0.82	2.14
Cronbach Alpha Person Raw Score Test Reliability = 0.80			

According to Bond & Fox (2001), a reliability coefficient of more than 0.8 is acceptable and has a strong value. The real person reliability and model person reliability coefficients of 0.80 and 0.82, respectively; and the item model reliability and item real reliability coefficients are both 0.92. Meanwhile, in terms of the separation index, person separation indicates the number of strata capabilities identified and item separation shows the separation of item difficulty levels (Ariffin, Omar, Isa, & Sharif, 2010). Low person separation implies that the instrument developed is not sensitive enough to differentiate the abilities of the sample so that it can be dealt with by adding more items while low item separation that the person sample is not large enough to confirm the hierarchy of item difficulty levels. As seen in Table 3, the values for both person separation and item separation show that the people sample has diverse abilities (high, medium, and low), and the level of item difficulty is well distributed and is on a logit scale with good reliability. Both values are by Linacre's (2023) recommendation that a good separation index is > 2.

Table 4 presents the item statistical values of item fit and item difficulty (*b*). From a statistical modeling perspective, item fit refers to whether a test item fits the test (Hayat, Putra, & Suryadi, 2020). For the Rasch Model, the criterion that can be used for item fit detection is the OUTFIT MNSQ (outlier-sensitive fit statistic-mean square), where statistics between 0.5 and 1.5 are considered productive of measurement (Linacre, 2023). Based on these criteria, it can be concluded that all items fit the Rasch model. Meanwhile, the item difficulty level describes the latent trait level (θ) where there is a 50% chance of a positive response from the item; for example, if $\delta = 0.75$, there is a 0.5 chance that someone with a latent trait level of 0.75 will respond positively to the item (Brown, 2015). In terms of the distribution of item difficulty, the values spread between -1.41 logits (item number 9) and 0.99 logits (item number 3). Item difficulty values usually range from -2.5 to +2.5 logits (Meijer & Tendeiro, 2018), which are further categorized as follows: easy item range $-2.5 \leq b < -1.0$; medium item range $-1.0 \leq b < +1.0$; and difficult item range $1.0 \leq$

b ≤ +2.5. Based on these criteria, it can be concluded that 90% of the items that make up the *Kalam* Cognitive Test instrument are in the medium difficulty category, while only 10% are included in the easy items (item number 1).

Table 4. Psychometric Characteristics

Item Number	OUTFIT MNSQ	b
1	0.81	-1.22
2	1.25	0.79
3	1.28	0.99
4	1.14	0.39
5	0.89	-0.75
6	0.77	-0.31
7	1.04	0.64
8	0.80	0.64
9	0.81	-1.41
10	1.15	0.25

The final psychometric characteristic under scrutiny pertains to the Andrich Rating Scale categorization, which is thoughtfully presented in Table 5. This categorization is an essential component of the assessment process, as it provides a structured framework for evaluating respondents' abilities within the context of the *Kalam* Cognitive Test. The comprehensive categorization scheme aids in precisely assessing the diverse levels of competence exhibited by respondents, offering a clear and systematic means of interpretation. In essence, this aspect of the analysis serves as a critical foundation for understanding how respondents' performance aligns with the established rating scale, ultimately contributing to the overall robustness and reliability of the assessment instrument.

Table 5. Summary of Category Structure

Category	Observed Count (%)	Observed Average	OUTFIT MNSQ	Andrich Threshold	Thresholds between categories (width)
2	34 (8)	-0.96	0.98	None	
3	130 (30)	-0.09	1.23	-1.94	-1.94
4	176 (40)	0.63	0.95	-0.03	-1,91
5	100 (23)	2.19	0.82	1,96	1,93

Five distinct criteria serve as the guiding benchmarks for interpreting the results presented in Table 5. Firstly, it is imperative to note that each category surpasses the minimum threshold of 10 observations, thereby satisfying the initial criterion. This comprehensive assessment encompasses all categories, where the observed frequencies span from 34 (8%) to 100 (23%). Notably, category one boasts the lowest observation count, while category four exhibits the highest. This fulfillment of the observation criterion underscores the reliability and robustness of the data collection process. It substantiates the notion that the frequency of observations within each category can be relied upon as a dependable metric for estimating the stability and consistency of the rating scale employed in the assessment process.

Secondly, a crucial criterion in our analysis involves the observation of a monotonically increasing trend in the average values as we move across categories (-0.96 < -0.09 < 0.63 < 2.19). To clarify, this means that as we progress from one category to the next, there is a consistent upward trajectory in the observed average values. For instance, in category 1, the observed average stands at -0.96 logits. This value signifies that the average estimated ability of all respondents who selected category 1 in any item of the *Kalam* Cognitive Test instrument is also -0.96 logits. This trend continues to hold as we move through the response categories, with the values increasing incrementally. This pattern is indicative of the fact that respondents with higher levels of ability tend to favor higher response categories, while those with lower abilities opt for lower categories. This observation aligns with the principles outlined in (Bond & Fox, 2015), further substantiating the credibility of our analysis.

Thirdly, it's crucial to highlight that the OUTFIT MNSQ values, all falling below the threshold of 2.0, reinforce the credibility of our findings. These values, ranging from 0.82 to 1.23 across all categories, firmly meet the stipulated criteria. This unequivocally affirms that undue variability or noise is absent within these categories, thereby validating their substantive significance in the measurement process. Furthermore, the fourth criterion pertains to the step calibration on the Andrich threshold, which consistently demonstrates a monotonic increase in correspondence with category increments. This observation underscores the alignment between the rating scale's steps and the underlying trait being measured, further bolstering the reliability of the *Kalam* Cognitive Test. Lastly, the fifth criterion, which evaluates the width of step calibration on the Andrich threshold within the range of -1.94 to 1.93 logits, adheres impeccably to the set standards. This conformance solidifies the assessment's precision and ensures that each step on the rating scale effectively differentiates between varying levels of competency. In culmination, all five criteria have been meticulously met, collectively reinforcing the robust psychometric properties of the *Kalam* Cognitive Test and endorsing its validity as a reliable assessment instrument.

The findings of this research hold significant implications in the realm of Islamic education. By conducting a thorough analysis of the psychometric characteristics of the *Kalam* Cognitive Test (KCT), this research underscores the importance of having a high-quality assessment instrument to measure the competence of prospective Islamic teachers in the field of *Kalam*. This can greatly assist educational institutions in improving the teaching and learning processes within this subject. The psychometric analysis results highlighted in the article emphasize the crucial need for ensuring that the assessment tool used to evaluate the abilities of future Islamic teachers possesses robust psychometric qualities. This assurance is vital for maintaining the validity and reliability of assessment outcomes. The article's reference to the fulfillment of criteria related to the observation frequency indicates the reliability and stability of the rating scale used in the KCT analysis. This insight can be applied in various educational and assessment contexts to ensure the consistent and dependable measurement of skills and abilities.

Therefore, this research contributes to the advancement of research in Islamic education by delving into the psychometric evaluation of a specific assessment tool, shedding light on the unique challenges and requirements within this field. The utilization of the Andrich Rating Scale model for analysis sets a methodological precedent. It contributes to the broader field of educational assessment by showcasing how advanced statistical techniques can be applied to enhance the quality of assessment tools. The research's findings and insights can be directly applied to improve the quality of teacher preparation programs for Islamic educators. By ensuring that the KCT effectively measures the competence of prospective Islamic teachers, it contributes to the preparation of better-qualified educators in the field of *Kalam*.

In summary, the article not only addresses a specific need in Islamic education but also provides methodological insights that can be applied in educational assessment across various disciplines. Its contributions extend to both the enhancement of educational practices and the advancement of research in the domain of Islamic studies.

CONCLUSION

In summary, this study's core objective, which was to thoroughly scrutinize the psychometric characteristics of the *Kalam* Cognitive Test through the utilization of the Andrich Rating Scale, has been successfully accomplished. The outcomes and insights derived from the meticulous analysis conducted in this research affirm that the study's initial goals have been effectively met. The unequivocal recommendation that arises from this study is the high suitability and efficacy of the Andrich Rating Scale as an indispensable tool for executing rating scale analyses within the realm of psychometrics. Its demonstrated ability to capture and represent the fundamental attributes of a robust assessment instrument underscores its significance in the evaluation and validation of such tools. Furthermore, the meticulous scrutiny of the empirical evidence gleaned from the analysis of the five criteria governing the functionality of the rating scale embedded within the *Kalam* Cognitive Test instrument unequivocally validates its appropriateness and reliability. This evidence attests to the soundness of the rating scale's design and its consistent performance in accurately gauging the abilities and competencies of prospective Islamic teachers in the domain of *Kalam*. The findings from this study hold substantial implications for the field of Islamic education, where the need for precise and trustworthy assessment tools is paramount, ensuring that future educators are adequately prepared to impart knowledge in this crucial subject. Consequently, the insights and recommendations emanating from this research serve as valuable contributions to the broader landscape of educational assessment, particularly within the context of Islamic studies, where such rigorous examinations of assessment tools are essential for maintaining the quality and integrity of education.

REFERENCES

Ajeigbe, T. O., & Afolabi, E. R. I. (2017). Assessing unidimensionality and differential item functioning in qualifying examination for Senior Secondary School Students, Osun State, Nigeria. *World Journal of Education*, *4*(4), 30-37. https://eric.ed.gov/?id=EJ1158579

Andersen, E. B. (1997). The rating scale model. In W. J. van der Linden, and R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 67-84). New York: Springer.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-573. https://doi.org/10.1007/BF02293814

Ariffin, S. R., Omar, B., Isa, A., & Sharif, S. (2010). Validity and reliability of multiple intelligent items using the Rasch measurement model. *Procedia-Social and Behavioral Sciences*, *9*, 729-733. https://doi.org/10.1016/j.sbspro.2010.12.225

Aziz, A.A., Masodi, M.S., & Zaharim, A. (2013). Asas model pengukuran rasch: pembentukan skala & struktur pengukuran. Bangi: Universiti Kebangsaan Malaysia

Baek, S.G. (2003). Measurement and assessment in teaching. *Asia Pacific Education Review*, *4*(2), 210-211.

Bakar, N. S. A., Maat, S. M., & Rosli, R. (2019). Evaluation of mathematics teachers' Technological Pedagogical Content Knowledge (TPACK) scale using Rasch model Analysis. *Religación: Revista de Ciencias Sociales y Humanidades*, *4*(16), 342-348. https://dialnet.unirioja.es/servlet/articulo?codigo=8274073

Bond, T.G. & Fox, C.M. (2001). *Applying the Rasch model fundamental measurement in the human sciences*. London: ERL Lawrence Baum Associates Publishers.

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

Clark, D. A., Donnellan, M. B., Durbin, C. E., Brooker, R. J., Neppl, T. K., Gunnar, M., Carlson, S. M., Le Mare, L., Kochanska, G., Fisher, P. A., Leve, L. D., Rothbart, M. K., & Putnam, S. P. (2020). Using item response theory to evaluate the Children's Behavior Questionnaire: Considerations of general functioning and assessment length. *Psychological Assessment, 32*(10), 928–942. https://doi.org/10.1037/pas0000883

Cordier, R., Speyer, R., Schindler, A., Michou, E., Heijnen, B. J., Baijens, L., Karaduman, A., Swan, J., Clave, P., & Joosten, A. V. (2018). Using Rasch analysis to evaluate the reliability and validity of the Swallowing Quality of Life Questionnaire: an item response theory approach. *Dysphagia*, *33*, 441-456. https://doi.org/10.1007/s00455-017-9873-4

DeMars, C. (2010). *Item response theory: understanding statistics measurement.* New York: Oxford University Press, Inc.

Gupta, C., Jain, A., & D'souza, A. S. (2016). Essay versus multiple-choice: A perspective from the undergraduate student point of view with its implications for examination. *Gazi Medical Journal*, *27*(1). https://medicaljournal.gazi.edu.tr/index.php/GMJ/article/view/1198

Hasbi, M. (2015). *Ilmu Kalam memotret berbagai aliran teologi dalam islam*. Trustmedia publishing: Yogyakarta.

Hayat, B., Putra, M. D. K., & Suryadi, B. (2020). Comparing item parameter estimates and fit statistics of the Rasch model from three different traditions. *Jurnal Penelitian dan Evaluasi Pendidikan*, *24*(1), 39-50. https://doi.org/10.21831/pep.v24i1. 29871

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational research review*, *2*(2), 130-144. https://doi.org/10.1016/j.edurev.2007.05.002

Lang, J. W., & Tay, L. (2021). The science and practice of item response theory in organizations. *Annual Review of Organizational Psychology and Organizational Behavior*, *8*, 311-338. https://doi.org/10.1146/annurev-orgpsych-012420-061705

Linacre, J.M. (2023). A user's guide to Winstep Ministeps: Rasch-Model computer programs. Chicago, IL.

Martinez, K. (1997). *The effect of a rubric on evaluating and improving student writing* (Doctoral dissertation, Caldwell College).

Meijer, R.R., & Tendeiro, J.N. 2018. Unidimensional item response theory. In P. Irwing, T. Booth, & D. J. Hugh (Eds.). *The Wiley handbook of psychometric testing : A multidisciplinary reference on survey, scale and test development* (pp. 413-433). Wiley. https://doi.org/10.1002/9781118489772.ch15

Minbashian, A., Huon, G.F., & Bird, K.D. 2004. Approaches to studying and academic performance in short-essay exams. *Higher Education*, 47, 161–176. https://doi.org/10.1023/B:HIGH.0000016443 43594.d1

Rusch, T., Lowry, P. B., Mair, P., & Treiblmaier, H. (2017). Breaking free from the limitations of classical test theory: Developing and measuring information systems scales using item response theory. *Information & Management*, *54*(2), 189-203. https://doi.org/10.1016/j.im.2016.06.005

Segal, D.L., & Coolidge, F.L. (2018). Reliability. In Bornstein, M. H. (Ed.). *The SAGE encyclopedia of lifespan human development* (pp 1835). Thousand Oaks, CA: SAGE Publications.

Sick, J. (2010). Assumptions and requirements of Rasch measurement. *Shiken: JALT Testing & Evaluation SIG Newsletter*, *14*(2), 23-29.

Sumekto, D. R., & Setyawati, H. (2018). Students descriptive writing performance: the analytic scoring assessment usage. *Cakrawala Pendidikan*, *37*(3), 413-425. https://doi.org/10.21831/cp.v38i3.20033

Tozoglu, D., Tozoglu, M.D., Gurses, A., & Dogar, C. (2004). The students' perceptions: Essay versus multiple-choice type exams. *Journal of Baltic Science Education*, 2(6), 52-59.

Tuckman, B. W. (1993). The essay test: A look at the advantages and disadvantages. *Nassp Bulletin*, *77*(555), 20-26. https://doi.org/10.1177/019263659307755504

Wahyuni, L. D., Gumela, G., & Maulana, H. (2021, June). Interrater Reliability: Comparison of essay tests and scoring rubrics. In *Journal of Physics: Conference Series* (Vol. 1933, No. 1, p. 012081). IOP Publishing.

Youssef, A. M. I. (2022). Using The Andrich Rating Scale Model (ARSM) to build a Scale for the Academic Proficiency Among Cairo University Students Psychometric Study. *Egyptian Journals,* *32*(16), 383-440. https://doi.org/10.21608/EJCJ.2022.247911

Zile-Tamsen, C.V. (2017). Using Rasch analysis to inform rating scale development. *Res High Educ*, *58*, 922-933. https://doi.org/10.1007/s11162-017-9448-0