SCIENTIFIC CULTURE
www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 12, No. 4, (2026), pp. 6503-6510
Open Access. Online & Print

DOI: 10.5281/zenodo.12426674

RELIABILITY ANALYSIS IN SCIENCE LITERACY LEARNING ENVIRONMENT INSTRUMENTS: GENERALIZABILITY THEORY

Siska Merrydian[1]*, Wardani Rahayu[2], Riyadi[3]

[1,2,3]Educational Research and Evaluation Study Program, Postgraduate Program of Jakarta State University, Indonesia.
Email: [1]siska.merrydian@mhs.unj.ac.id, [2]wardani.rahayu@unj.ac.id, [3]riyadi@unj.ac.id

Received: 07/11/2025
Accepted: 09/01/2026

Corresponding Author: Siska Merrydian (siska.merrydian@mhs.unj.ac.id)

ABSTRACT

The reliability of science literacy learning environment instruments presents its own challenges because of the dependence on person and items. This study applied the reliability of McDonald's Omega (ω) and Generalization Theory (GT) to evaluate the reliability of science literacy learning environment instruments in madrasah aliyah. McDonald's Omega reliability (ω) calculated, followed by GT 2 (two) facet (person x item) design involving 923 person and 43 items. The reliability results of McDonald's Omega (ω) show a very high category. The results of the Generalizability Theory analysis through the G Study showed that the largest variance came from person-item interaction (P×I) of 0.936 or 60.8%, which indicated an inconsistency in the person's response to the item. The generalization coefficient of both relative and absolute is 0.97, which indicates the reliability of the instrument is very high. The results of the D Study show that with 43 items, the instrument has reached an optimal level of reliability, adding and reducing items does not provide a significant increase in reliability.

KEYWORDS: Learning Environment, Science Literacy, Mcdonald's Omega Reliability, Generalizability Theory.

INTRODUCTION

Reliability is defined as the consistency of the results obtained from a test. According to Dragon (2022), reliability reflects the level of trust in the data. Sumintono & Widhiarso (2015) mentioning that reliability shows the extent to which measurements made repeatedly produce consistent information. To ensure reliable measurement results, the instrument needs to be tested to obtain an estimate of reliability. It is concluded that reliability can be understood as the level of consistency and trust in the results of a measurement. An instrument is said to be reliable if it is able to produce stable and similar data when used repeatedly under comparable conditions.

Reliability is one of the main sources of evidence in determining the consistency produced from test scores and measurement results (Kane, 2006; Zumbo, 2007). The term test here includes various forms of quantitative measurement tools with many items, such as tests, scales, and surveys that are commonly used in social science research.

Mcdonald's Omega reliability (ω) is an internal reliability coefficient based on factor analysis used to measure how much of the proportion of variance in an instrument's total score actually reflects the latent construct being measured. Mcdonald's Omega (ω) is calculated from the load factor and the variance error results of factor analysis, thus taking into account the difference in the strength of each item in representing the factor. The value ω indicates a more realistic level of internal consistency and is highly recommended for modern psychometric scales because it provides a more accurate estimate of reliability on instruments with non-uniform item weights.

The use of McDonald's Omega reliability (ω) relevant in the context of IRT analysis because this approach is based on a latent model that takes into account the difference in the contribution of each item to the constructed being measured. In IRT, each item has a discrimination parameter that shows the strength of its relationship to the latent trait. McDonald's Omega (ω) is calculated based on factor model parameters that are in line with the basic principle of IRT, so that it is able to reflect variations in the characteristics of the item more accurately. McDonald's Omega (ω) provides reliability estimates consistent with the model-based measurement framework used in IRT analysis.

Reliability coefficient Mcdonald's Omega (ω) represents a true estimate of population reliability by eliminating certain items from the scale. Reliability calculations Mcdonald's Omega (ω) in line with the confidence interval pay attention to variations in the estimation process in more detail, providing a more precise level of validity regarding the consistency of the scale structure (Ravinder & Saraswathi, 2020). Some studies that use reliability Mcdonald's Omega (ω), research W. Rahayu & Harjono (2015); Deng & Chan (2017); Ravinder & Saraswathi (2020); Paola *et al*. (2023); and Orcans (2023).

Generalization theory is seen as a comprehensive psychometric approach to dealing with the complexity of measurement on assessments that are based on local contexts (Zhang, 2006). A number of studies have documented its use in various fields. In performance assessments involving raters, generalization theory is able to map the sources of variance in detail and shows that the greatest contribution often comes from interactions between person and items, while differences between raters tend to be smaller (Peeters) *et al*., 2021). Through its ability to identify and map various sources of variance, this approach shows that the greatest variation in performance appraisals generally comes from interactions between person and items, while the influence of differences between raters is relatively small. Generalization theory provides a solid basis for improving the accuracy and quality of measurement results.

In language assessment, generalization theory is used to examine the factors that affect writing scores and find that the application of analytical assessments, integrated writing tasks, and the involvement of at least four raters can result in high reliability and generalization power. In fact, a small group of trained assessors can still achieve an adequate level of reliability (Lertsakulbunlue & Kantiwong, 2025).

In the validation process by experts, generalization theory analysis generally shows that the variation in scores is more due to differences between items, while the variation between assessors is relatively low (Dorathy *et al*., 2021). In addition, the application of generalization theory also extends to educational and psychological evaluation, the development of career interest instruments, psychological inventories, and medical education, which further emphasizes its flexibility and contribution in improving the quality of measurement (Clayson *et al.*, 2021). Generalization theory suggests that in expert validation, score variations are influenced more by item differences than by raters. Its wide application in various fields confirms its versatility and role in improving measurement quality.

Generalization theory provides a comprehensive approach to assessing the reliability of instruments in

complex situations. Generalization theory separates score variance into various sources, such as differences between items, between person, and interactions between the two. With more detailed mapping of error sources, it is possible to understand more deeply the factors that affect the consistency of measurement results (Brennan, 2001). This approach is particularly relevant for instruments developed in local contexts, as variations in items and interpretations by individuals can affect the instrument's reliability.

The advantage of this research lies in its novelty in integrating reliability measurement using McDonald's Omega coefficient (ω) and Generalizability Theory to evaluate the instruments of science literacy learning environment in madrasah aliyah. Until now, studies that simultaneously utilize these two approaches in the context of madrasah education have not existed, even though the combination of the two is able to provide a more comprehensive picture of the internal consistency of the instrument as well as the sources of variation in measurement errors. This research not only strengthens the empirical reliability of the instrument, but also makes a methodological contribution to the development of more accurate and contextual learning.

METHODS

Participants

This study involved 923 grade 11 students at MAN 19 Jakarta, MAN 11 Jakarta, MAN 15 Jakarta, MAN 16 Jakarta, MAN 6 Jakarta, MAN 10 Jakarta, MAS Al Hamid, MAS Nurul Falah, MAS Annida Al-Islamy dan MAS Pembangunan UIN Jakarta, Indonesia.

Data Collection

The research was carried out in the even semester of 2024/2025 in June 2025 using a google form. Likert scales are used: 5 (always), 4 (often), 3 (sometimes), 2 (rarely), and 1 (never). Consideration for the selection of DKI Jakarta because it is an urban city with diverse populations from various regions.

Data Analysis

The reliability coefficient indicates the reliability of a set of corks, the value of which ranges from 0 to 1. The reliability coefficient of 1 indicates that the differences between the observed test sectors of the respondents are fully consistent with the differences between the true sectors. On the other hand, a reliability of 0 indicates that the differences between the observed test sectors are not at all consistent with the differences between the true sectors (Furr, 2022). Guilford (1965) categorize the reliability test into 4 categories: 0.80-1.00, very high reliability; 0.60-0.80, high reliability; 0.40-0.60, medium reliability, and 0.20-0.40, low reliability. According to Three & Quainoo (2019) The common interpretation of the reliability coefficient is < 0.5 for low reliability, 0.5 – 0.8 for medium reliability (acceptable), and > 0.8 for high reliability (good). The reliability value ranges from 0 to 1, where a higher number indicates that the items in the instrument consistently measure the same thing. Conversely, if the value is close to 0, it means that there are items that are out of alignment or do not measure the same construct.

Mcdonald's Omega reliability coefficient formula (ω) (Paola *et al*., 2023):

$$\omega = \frac{\left(\sum_{i=1}^{k} \lambda_i\right)^2}{\left(\sum_{i=1}^{k} \lambda_i\right)^2 + \sum_{i=1}^{k}\left(1 - \lambda_i{}^2\right)}$$

Description:

ω = Mcdonald's Omega reliability coefficient

k = number of grains

λ_i = Load factor item I

$1 - \lambda^2$ = Error Variance

Generalization theory is a development of classical test theory that aims to provide a more comprehensive estimation of measurement reliability by identifying various sources of error in a measurement design. In this framework, the G Study is used to estimate the variance components of various facets, such as person, items, and interactions, so that the contribution of each source to the variation in the total score can be known. The results of the G Study became the basis for evaluating the quality of the instruments and identifying the main sources of measurement errors. Furthermore, the D Study utilizes the variance component obtained from the G Study to simulate and evaluate various measurement design alternatives, such as changes in the number of items or other measurement conditions, in order to obtain an optimal level of reliability. The combination of the G Study and the D Study allows researchers not only to understand the structure of measurement errors, but also to design more reliable instruments according to the measurement objectives (Brennan, 2001).

RESULTS AND DISCUSSION

Descriptive Statistics

The science literacy learning environment (SLILEI) instrument in madrasah aliyah consists of 8 dimensions with a total of 43 items, namely: science communication (9 items), science investigation (3 items), use of scientific information (4 items), teacher

support (7 items), cooperation (3 items), involvement (5 items), justice (6 items), and science spirituality (6 items).

Based on table 1, all items have the same score range, namely a minimum score of 1 and a maximum of 5, which indicates that the entire scale is fully utilized by the person, in this case the student. The mean values in each aspect were relatively close to each aspect, ranging from 2.849 to 2.887, with the highest score in Science Investigation (IS) at 2.887 and the lowest in Cooperation (KJ) at 2.849. This indicates that a person's perception of each dimension tends to be at a moderate level. The standard deviation values in the range of 1.231 to 1.259 indicate a fairly even variation in response across each item, with the largest variation in Science Investigation (IS) and the smallest variation in Scientific Information Use (PI) and Teacher Support (DG). Overall, this data reflects a relatively consistent distribution of responses across all aspects measured.

Table 1. Descriptive Statistics.

Item	Minimum	Maximum	Mean	Std. Deviation
Science Communication (KS)	1	5	2.867	1.243
Science Investigation (IS)	1	5	2.887	1.259
Use of Scientific Information (PI)	1	5	2.874	1.231
Teacher Support (DG)	1	5	2.861	1.231
Cooperation (KJ)	1	5	2.849	1.232
Involvement (KT)	1	5	2.863	1.241
Justice (KD)	1	5	2.860	1.251
Spirituality of Science (SS)	1	5	2.863	1.240

Mcdonald's Omega (ω)

The results of the reliability calculation presented are Mcdonald's Omega (ω). The categorization of reliability results according to Guilford (1965). The results of the reliability calculation of Mcdonald's Omega (ω) can be seen in table 2.

Table 2. Mcdonald's Omega (ω) Reliability.

Dimensions	Mcdonald's Omega (ω)	Remarks
Science Communication (KS)	0.945	Very High
Science Investigation (IS)	0.826	Very High
Use of Scientific Information (PI)	0.877	Very High
Teacher Support (DG)	0.917	Very High
Cooperation (KJ)	0.838	Very High
Involvement (KT)	0.867	Very High
Justice (KD)	0.921	Very High
Spirituality of Science (SS)	0.898	Very High
SLILEI	0.965	Very High

Based on table 2. the Science Communication (KS) dimension has a reliability value of Mcdonald's Omega (ω) of 0.945 which is included in the very high category. This value shows that the instrument in the science communication dimension has excellent internal consistency in measuring the construct in question. The Science Investigation Dimension (IS) obtained a Mcdonald's Omega (ω) score of 0.826 in the very high category. This indicates that the items in this dimension are quite consistent in representing the investigative capabilities of science, although their value is lower than some of the other dimensions.

In the Scientific Information Use (PI) dimension, the value of Mcdonald's Omega (ω) was recorded at 0.877 which is also in the very high category. This shows that the instrument is able to measure the ability to use scientific information consistently. The Teacher Support (DG) dimension has a Mcdonald's Omega (ω) value of 0.917 with a very high category. This score reflects that perceptions of teacher support are measured by an excellent level of consistency.

Furthermore, the Cooperation dimension (KJ) shows a Mcdonald's Omega value (ω) of 0.838 which is in the very high category. This indicates that instruments in the cooperation dimension are reliable in capturing aspects of collaboration between student. The Involvement Dimension (KT) obtained a Mcdonald's Omega (ω) value of 0.867 in the very high category. This shows that the level of person involvement can be measured consistently through the instruments used.

In the Justice (KD) dimension, the value of Mcdonald's Omega (ω) reached 0.921 which was in the very high category. This value shows that the instrument is very consistent in measuring the perception of justice. The Dimension of Science Spirituality (SS) has a Mcdonald's Omega (ω) value of 0.898 in the very high category. This shows that aspects of spirituality in science are measured by a strong level of reliability.

Overall, the science literacy learning environment instrument has a Mcdonald's Omega (ω) value of 0.965 which is in the very high category. This value confirms that the instrument in general has a very strong internal consistency and is reliable in measuring all dimensions studied.

Table 3. Observation and Estimation Designs.

Facet	Label	Levels	Univ.
Person	P	923	INF
Item	I	43	INF

Table 4. Analysis of Variance.

				Components				
Source	**SS**	**df**	**MS**	**Random**	**Mixed**	**Corrected**	**%**	**SE**
P	24803.242	922	26.902	0.604	0.604	0.604	39.2	0.029
I	25.184	42	0.599	0.000	0.000	0.000	0.0	0.000
PI	36253.003	38724	0.936	0.936	0.936	0.936	60.8	0.007
Total	**61081.428**	**39688**					**100%**	

Table 5. G Study Table.

(Measurement Design P/I)

Source or Variance	**Differenti-ation Variance**	**Source or Variance**	**Relative Error Variance**	**% Relative**	**Absolute Error Variance**	**% Absolute**
P	0.604					
		I			0.000	0.000
		PI	0.022	100.0	0.022	100.0
Sum of Variances	0.604		0.022	100%	0.022	100%
Standard Deviation	0.777		Relative SE: 0.148		Absolute SE: 0.148	
Coef_G Relative 0.97						
Coef_G Absolute 0.97						

Grand mean for level used: 2.865
Variance error of the mean for levels used: 0.000
Standard error of the grand mean: 0.026

G Study

Based on tables 4 and 5, in the Person × Item (P × I) measurement design information was obtained about the sources of score variations. The variance derived from person (P) was 0.604 with a contribution of 39.2% of the total variance. This shows that the difference in scores between students is quite significant and reflects a variation in student behavioral tendencies, which is indeed the main goal in a measurement.

Furthermore the variance derived from item (I) is 0.000. This value indicates that there is no significant difference in difficulty between items. so that all items can be said to have relatively homogeneous characteristics.

The largest source of variation comes from the interaction between person and item (P×I) with a variance value of 0.936 or 60.8% of the total variance. This component reflects a combination of student-specific interactions with specific items and random errors. The magnitude of P×I contribution shows that student responses tend to be inconsistent between items, thus becoming the main source of measurement errors.

Based on the results of the analysis, the source that most affected the variation in scores in this study was the interaction between person and item (P×I) with a variance value of 0.936 or contributing 60.8% of the total variance. This shows that the difference in scores that appear is more dominant due to the inconsistency of students' responses to different items not solely due to differences in students behavioral tendencies or the difficulty level of the item. A student can show different performance on each item. So this interaction is a major source of variation as well as measurement errors. Meanwhile, the variance derived from the person factor was only 39.2%, which still shows a difference in behavioral tendencies between students and the variance of items that is close to zero indicates that all items have a relatively uniform level of difficulty. It can be concluded that the interaction factor between person and item is the biggest contributor in influencing the score results in this study.

In the G Study table, the differentiation variance comes only from the person component (0.604), which shows that the relevant variation to distinguish students does lie in the differences between students. Meanwhile, the relative error variance and absolute error variance are both 0.022 which are all derived from the P×I interaction component.

The standard error value of both relative and absolute is 0.148 which is relatively small. This indicates that the error rate in the measurement is relatively low. A G-coefficient value of both relative and absolute is 0.97. This value indicates that the instrument has a very high level of reliability, both for the purpose of comparison between students and for absolute decision-making.

D Study

Based on the results of the D Study (see tables 4 and 5), the analysis was focused on evaluating the effectiveness of the measurement design used namely with 923 person and 43 items. By using the variance component of the G Study, it was obtained

that the variance error was 0.022 which resulted in a G-coefficient value of 0.97.

These results show that with the number of items in use today, the instrument has achieved a very optimal level of reliability. In the context of the D Study. the number of items is an important factor that affects the magnitude of the error variance. The more items used, the more error variance tends to decrease so reliability increases.

However, in this case, since the reliability value is already high, the increase in the number of items is not expected to provide a significant increase. Furthermore, because the variance of items in the G Study was zero, the change in the number of items in the D Study had more effect on reducing errors than improving the quality of the items themselves. The quality of items is relatively good and uniform. So design optimization is more related to the efficiency of the number of items.

The results of Study D indicate that the measurement design used is highly reliable. The instrument can be used optimally under its current conditions with 43 items, with the addition or removal of items not yielding significant increases in reliability.

CONCLUSION

The results of the reliability analysis using the Mcdonald's Omega (ω) coefficient showed that the instrument had a very high level of internal consistency. Indicating that all items were able to measure the same construct stably. These findings are in line with the results of the Generalizability Theory in the G Study, where the generalization coefficient reached 0.97 for both relative and absolute decisions. Indicating strong instrument reliability. The G Study also showed that the largest source of variation came from person-item interaction (P×I), which reflects the inconsistency of students' responses to certain items and is the main source of measurement errors. The results of the D Study confirm that with a total of 43 items, the instrument has reached an optimal level of reliability. So the addition or subtraction of items does not provide a significant improvement. The combination of Mcdonald's Omega's (ω) reliability results and the G and D Study analyses showed that the instrument was not only internally consistent, but also stable under various measurement conditions.

ACKNOWLEDGMENTS

The author expresses his highest gratitude and appreciation to the Indonesian Ministry of Higher Education, Science and Technology (Kemendiktisaintek RI) as the grantor of doctoral dissertation research with master contract number 083/C3/DT.05.00/PL/2025 dated May 28, 2025 and derivative contract number 9/UN39.14/C3/DT.05.00/PPS-PDD/PL/2025 dated June 03, 2025.

REFERENCES

Clayson. P. E.. Carbine. K. A.. Baldwin. S. A.. Olsen. J. A.. & Larson. M. J. (2021). Using generalizability theory and the ERP Reliability Analysis (ERA) Toolbox for assessing test-retest reliability of ERP scores part 1 : Algorithms . framework . and implementation. *International Journal of Psychophysiology*. *December 2020*. https://doi.org/10.1016/j.ijpsycho.2021.01.006

Deng. L.. & Chan. W. (2017). *Testing the Difference Between Reliability Coefficients Alpha and Omega*. https://doi.org/10.1177/0013164416658325

Three. S. O.. & Quainoo. H. (2019). *Reliability of assessments in engineering education using Cronbach ' s alpha . KR and split-half methods*. *21*(1). 24–29.

Furr. R. M. (2022). Psychometrics: An intoduction (Fourth). In *SAGE Publications Ltd*.

Guilford. J. P. (1965). Fundamental Statistics in Psychology and Education. In *New York: McGraw-Hill*.

Kane. M. N. (2006). *Educational Measurement* (4th ed). American Council on Education and National Council on Measurement in Education.

Lertsakulbunlue. S.. & Kantiwong. A. (2025). Evaluating the dependability of peer assessment in project - based learning for pre - clinical students : a generalizability theory approach. *BMC Medical Education*. https://doi.org/10.1186/s12909-025-06772-0

Naga. D. S. (2022). *Sekor Theory on Mental Measurement*. A Taste of Cincinnati.

Orcan. F. (2023). *Comparison of cronbach ' s alpha and McDonald ' s omega for ordinal data : Are they different ? 10*(4). 709–722.

Paola. C.. Schwall. P.. Meesters. C.. & Hardt. J. (2023). Social Sciences & Humanities Open Estimating reliability : A comparison of Cronbach ' s α . McDonald's ω t and the greatest lower bound. *Social Sciences & Humanities Open*. *7*(1). 100368. https://doi.org/10.1016/j.ssaho.2022.100368

Peeters. M. J.. Cor. M. K.. Petite. S. E.. & Schroeder. M. N. (2021). *Note Validation Evidence using Generalizability Theory for an Objective Structured Clinical Examination*. *12*(1). 1–5.

Rahayu. W.. & Harjono. D. (2015). *Estimation Comparison of Multidimensional Reliability Coefficients Measurement of Senior High School Students' Affection towards Mathematics*. *3*(11). 1444–1449. https://doi.org/10.12691/education-3-11-15

Ravinder. E. B.. & Saraswathi. A. B. (2020). *Literature Review Of Cronbachalphacoefficient (A) And McDonald's Omega Coefficient (Ω)*. *7*(06). 2943–2949.

Robert L. Brennan. (2001). *Generalizability Theory*. University of Iowa.

Sumintono. B.. & Widhiarso. W. (2015). *The Application of Rasch Modeling in Educational Assessment* (October Issue). Trim Communicate.

Sunday Dorathy. Dr. A. Amadioha. D. G. W. O. (2021). *Application of Generalizability Theory in the Estimation Dependability of Critical*. *8056*(9). 171–178. https://doi.org/10.36347/sjpms.2021.v08i09.002

Zhang. S. (2006). *Investigating the relative effects of persons . items . sections . and languages on TOEIC score dependability*. *23*(1999). 351–369.

Zumbo. B. D. (2007). *Handbook of Statistics*. The Netherlands: Elsevier Science B.V.

APPENDIX

SLILEI Instruments

	Science Communication
KS1	I convey the results of my thoughts orally with the support of accurate evidence
KS2	I convey the results of my thoughts orally with logical explanations
KS3	I conveyed verbally how scientific knowledge can be applied in daily life
KS4	I write evidence that supports the scientific research hypothesis using reliable references
KS5	I write solutions to scientific problems without paying attention to the regularity of the existing data
KS6	I wrote about a research experiment model based on relevant theories
KS7	I wrote a research hypothesis about natural phenomena using a systematic structure
KS8	I present the results of scientific research into the form of interesting pictures
KS9	I present the results of scientific investigations in the form of an easy-to-understand table
	Science Investigation
IS1	I propose measures in scientific research
IS2	I identify sources of errors that may affect the results of scientific investigations
IS3	I ignore possible errors in the results of scientific research
	Use of Scientific Information
PI1	I make use of valid information before drawing conclusions
PI2	I always filter information before taking action
PI3	I give an opinion based on accurate evidence
PI4	I am responding to a less logical scientific argument
	Teacher Support
DG1	Teacher presents informative video of scientific phenomena
DG2	The teacher invited me to do a practicum at school
DG3	The teacher took me to the library to collect learning materials
DG4	Teachers give examples of how to identify natural phenomena around the school
DG5	Master ignored my opportunity to conduct scientific investigation
DG6	The teacher guided me in completing a scientific investigation project
DG7	The teacher gave me the opportunity to present the results of scientific investigations
	Collaboration
KJ1	I work with friends to complete project tasks
KJ2	I share ideas with friends to improve my understanding of learning materials
KJ3	I help a friend who is having difficulty understanding the material
	Involvement
KT1	I pay serious attention when a friend expresses an opinion
KT2	I give positive affirmations in group discussions
KT3	I actively ask questions during group discussions
KT4	I feel comfortable when I read the scientific articles in the library
KT5	I am motivated to learn because teachers use interactive media in learning
	Justice
KD1	I got the same opportunity using practicum tools and materials
KD2	I was given adequate opportunity to study the learning tutorial videos
KD3	I was given the same opportunity by the teacher to do an experiment in the school yard
KD4	I conduct scientific research with direction from a teacher
KD5	I was given appreciation by the teacher when I did the assignment presentation
KD6	I was given the same opportunity by the teacher to convey ideas
	Spirituality of Science
SS1	I realize that science must be used for the benefit of mankind according to the teachings of Islam
SS2	I believe that natural phenomena and religion are at odds
SS3	I see the connection between natural phenomena and the signs of God's power
SS4	I associate science knowledge with the teachings of Islam in daily life
SS5	I did the experiment with full honesty as part of the teachings of Islam
SS6	I convey the results of scientific observations by paying attention to Islamic principles