Al-Ishlah: Jurnal Pendidikan
Vol.17, 4 (December, 2025), pp. 6963-6970
ISSN: 2087-9490 EISSN: 2597-940X, DOI: 10.35445/alishlah.v17i4. 8200

Generalizability Theory Analysis of Local Curriculum Validation Instruments: Item Effects and Rater Consistency

Ruslina Irianty[1], , Rustam A.R. Selang[2], Helli Ihsan[3], Yetty Supriyati[4], Ilham Falani[5]

[1] Universitas Negeri Jakarta, Jakarta Timur, Indonesia; ruslina.irianty@mhs.unj.ac.id
[2] Universitas Negeri Jakarta, Jakarta Timur, Indonesia; rus.selang85@gmail.com
[3] Universitas Pendidikan Indonesia, Bandung, Indonesia: helliihsan@upi.edu
[4] Universitas Negeri Jakarta, Jakarta Timur, Indonesia; yetti.supriyati@unj.ac.id
[5] Universitas Negeri Jakarta, Jakarta Timur, Indonesia; ilhamfalani@unja.ac.id

ARTICLE INFO

Keywords:

generalizability theory;
validation instrument;
inter-rater reliability;
local curriculum;
educational measurement

Article history:

Received 2025-07-04
Revised 2025-07-15
Accepted 2025-12-30

ABSTRACT

Validating assessment instruments for local curricula presents unique challenges due to context-specific content and reliance on expert judgment. This study applies Generalizability Theory (GT) to evaluate the reliability of a curriculum validation instrument designed for the Social Studies Local Plants program in Fakfak, Indonesia. A fully crossed GT design was used, involving 26 items evaluated by 3 expert validators, yielding 78 observations. Each expert rated all items using a 4-point Likert scale. Variance components were estimated using a linear mixed-effects model with REML, implemented via the *lme4* package in R. The analysis revealed that item variance accounted for 98.2% of total score variance (σ^2 = 291.97), indicating strong item discriminability. Rater variance (0.2%) and item × rater interaction (1.1%) were minimal, demonstrating high inter-rater consistency. The Generalizability Coefficient (G = 0.995) and Dependability Coefficient (Φ = 0.986) exceeded the thresholds for both relative and absolute decision-making. A D-study showed that high reliability ($\Phi \geq 0.90$) could be maintained with as few as 2 raters and 15 items. The instrument demonstrated excellent reliability and is suitable for evaluating local curriculum validity. Minimal rater-related variance suggests that future improvements should focus on item refinement rather than rater training. These findings support the broader use of GT in educational instrument validation, particularly in context-rich, expert-judged settings.

Corresponding Author :
Ruslina Irianty
Universitas Negeri Jakarta, Jakarta Timur, Indonesia; ruslina.irianty@mhs.unj.ac.id

1. INTRODUCTION

The development and implementation of local curricula, such as the Social Studies Local Plants program for junior high schools in Fakfak, Indonesia, present unique challenges for educational measurement. These curricula are designed to reflect regional culture, ecology, and values, such as the integration of nutmeg cultivation and the philosophy of "one stove three stones"—to strengthen students' identity and character. However, the contextual specificity of such curricula complicates the validation of

assessment instruments, as conventional psychometric approaches may not fully capture the nuances of local content and expert judgment.

Generalizability Theory (GT) offers a robust framework for evaluating the reliability of validation instruments in these complex settings. Unlike Classical Test Theory, which aggregates all error into a single undifferentiated term, GT decomposes score variance into multiple sources—such as items, raters, and their interactions—enabling a more nuanced understanding of measurement dependability. This is particularly valuable for locally contextualized rubrics, where both item content and rater interpretation may vary.

Despite the growing adoption of local curricula, there is limited psychometric evidence supporting the reliability and validity of their associated assessment rubrics. Specifically, few studies have systematically examined how well expert-validated instruments perform in capturing the intended constructs within a local context. This study addresses this gap by applying GT to evaluate the reliability of a 26-item validation instrument rated by three expert validators for the Fakfak local curriculum. The research questions are: (1) What are the main sources of score variance in the instrument? (2) How reliable are the expert ratings as measured by GT coefficients?

Several studies have demonstrated the utility of GT in similar contexts. For example, (Peeters et al., 2021) used GT to validate clinical examination instruments, finding that item variance dominated and rater consistency was high, paralleling our design and findings. (Lertsakulbunlue & Kantiwong, 2025)Applied GT to peer assessment in medical education, showing that a small number of well-trained raters can yield high reliability (Clayson et al., 2021). Used GT to assess the reliability of ERP scores, highlighting the importance of disentangling item and rater effects. Finally, (Dorathy et al., 2021) Estimating the dependability of a critical thinking scale for university students using GT, emphasizing the method's value in educational instrument validation. These studies collectively support the use of GT for evaluating complex, context-specific instruments and inform the methodological choices in this research. In summary, this study leverages Generalizability Theory to provide psychometric evidence for the reliability of a local curriculum validation instrument, with a focus on identifying key sources of measurement error and informing future instrument development.

The development and implementation of local curricula, such as the Social Studies Local Plants program in Fakfak, Indonesia, present unique challenges for educational measurement. These curricula are designed to reflect the unique cultural, ecological, and social values of a region, aiming to strengthen students' identity and character through the integration of local wisdom, such as nutmeg cultivation and the philosophy of "one stove three stones." However, the contextual specificity of these curricula complicates the process of validating assessment instruments, as conventional psychometric approaches may not fully capture the nuances of local content and expert judgment. Ensuring that such instruments are both reliable and valid is essential for supporting meaningful educational decisions and for building a strong validity argument for the curriculum as a whole.

Generalizability Theory (GT) has emerged as a comprehensive and powerful psychometric framework for addressing the complex measurement issues inherent in locally contextualized assessments. (Zhang, 2006). Unlike Classical Test Theory (CTT), which partitions observed scores into "true" scores and a single undifferentiated error term (Crocker & Algina, 1986)GT allows researchers to disentangle multiple sources of measurement error by decomposing score variance into components attributable to various facets, such as items, raters, tasks, and their interactions.(Brennan, 2000a). This multifaceted approach is particularly valuable for local curriculum validation, where both item content and rater interpretation may vary significantly. In GT, a facet is any characteristic of the measurement situation that can be varied, such as test items or raters, and the object of measurement—such as a student, class, or, as in this study, the items themselves—is the primary source of interest.

The GT analysis process typically involves two main stages. The first is the Generalizability Study (G-study), which estimates the variance components associated with different sources of variation in the universe of admissible observations. This stage helps researchers understand how much each source, such as a rater or item, contributes to the total variability in scores. The second stage, the Decision Study (D-

study), uses these variance component estimates to design and optimize future measurement procedures, such as determining the number of items or raters needed to achieve a desired level of reliability. GT distinguishes between two types of reliability coefficients: The Generalizability Coefficient (G-coefficient), which is relevant for relative decisions such as ranking, and the Dependability Coefficient (Phi-coefficient), which is used for absolute decisions such as pass/fail outcomes. The ability to estimate the average reliability of ratings for each participant, while accounting for inter-rater and intra-rater inconsistencies, makes GT especially suitable for complex validation contexts.

A key advantage of GT over CTT is its capacity to identify and estimate the magnitude of each source of error simultaneously, providing a richer and more multidimensional understanding of measurement error (Brennan, 2021). For example, while CTT assumes that increasing test length always increases reliability, GT reveals that this is not necessarily the case if multiple random facets contribute to error, including inter-rater and intra-rater inconsistencies. This nuanced perspective enables researchers to manipulate the number of raters, tasks, or items and to obtain reliability estimates that are comparable across different tests and contexts. GT thus empowers instrument developers to make informed decisions about the design and use of assessment tools, particularly in settings where local context and expert judgment play a critical role. (Brennan, 2000b).

Despite the growing adoption of local curricula, there remains limited psychometric evidence supporting the reliability and validity of their associated assessment rubrics. Few studies have systematically examined how well expert-validated instruments perform in capturing the intended constructs within a local context. This study addresses this gap by applying GT to evaluate the reliability of a 26-item validation instrument rated by three expert validators for the Fakfak local curriculum. The research focuses on identifying the main sources of score variance in the instrument and assessing the reliability of expert ratings as measured by GT coefficients. By understanding these sources of variation, the study aims to provide clear implications for improving the quality of instruments and validation processes in similar educational settings.

The application of GT in instrument validation has been demonstrated in a variety of contexts. For instance, studies have shown that GT is well suited to evaluating performance-based assessments involving human raters, where student-task interactions often contribute significantly to variance, while rater facets contribute relatively little. (Peeters et al., 2021). In language assessment, GT has been used to analyze factors affecting writing scores, revealing that peer rating, analytical assessment methods, and integrated writing tasks can yield highly reliable and generalizable results, especially when at least four raters are involved. Peer assessment studies have found that even a small number of well-trained raters can achieve high reliability coefficients. (Lertsakulbunlue & Kantiwong, 2025). In the context of expert validation, GT has been used to analyze the reliability of validated instruments, often finding that item variance dominates and inter-rater variance is minimal, indicating high consistency among experts. (Dorathy et al., 2021). GT has also been applied in educational and psychological evaluation, the development of career choice instruments, psychological inventories, and even in medical education, further demonstrating its versatility and value for improving measurement quality. (Clayson et al., 2021).

These studies collectively support the use of GT for evaluating complex, context-specific instruments and inform the methodological choices in this research. By leveraging GT, this study provides psychometric evidence for the reliability of a local curriculum validation instrument, with a focus on identifying key sources of measurement error and informing future instrument development. The findings are expected to contribute to the broader literature on educational measurement and to offer practical guidance for researchers and practitioners involved in the validation of locally contextualized assessment tools.

2. METHODS

Participants in this study were three expert validators selected based on their advanced qualifications in curriculum development, psychometrics, and local educational context. The criteria for expert selection included a minimum of five years' experience in educational assessment or curriculum

evaluation, prior involvement in local curriculum projects, and advanced academic credentials in education or related fields. The use of three experts is justified by prior simulation studies and empirical research in Generalizability Theory, which have shown that a small, well-trained panel can yield high reliability coefficients, especially when rater consistency is optimal (Lertsakulbunlue & Kantiwong, 2025). While increasing the number of raters can improve reliability, the marginal benefit diminishes after a certain point, and three experts are often sufficient for robust reliability estimation in fully crossed GT designs.

The instrument consisted of 26 items developed through a rigorous process involving literature review, expert consultation, and iterative refinement. Initial item generation was informed by a review of national and local curriculum standards, as well as prior validated instruments in similar contexts. Items were then reviewed and refined in collaboration with subject matter experts to ensure content validity and contextual relevance. Each item was operationally defined to measure a specific aspect of the curriculum's validity, such as alignment with local values, clarity of learning objectives, and appropriateness of assessment methods. For example, one item asked experts to rate the extent to which the curriculum integrates local ecological knowledge, while another focused on the clarity of competency statements. Sample items included: "The curriculum content reflects the unique cultural identity of the Fakfak region," and "Assessment methods are appropriate for the targeted learning outcomes."

The rating process employed a four-point Likert scale, ranging from 1 (not appropriate) to 4 (highly appropriate), to capture the degree of agreement with each item's statement. Before the rating session, experts participated in a calibration meeting to discuss the operational definitions of each item and to align their interpretations of the scale points. This training session was designed to minimize rater drift and ensure a shared understanding of the rating protocol. The rating process was conducted over a one-week period, during which each expert independently rated all 26 items. Ratings were submitted electronically, and no communication between raters was allowed during the scoring period to prevent bias or influence.

For data analysis, a linear mixed-effects model was specified as Score ~ (1 | Statement) + (1 | Rater) + (1 | Statement: Rater), following best practices in Generalizability Theory for fully crossed designs (Brennan, 2000c; Cetin et al., 2016). The inclusion of the Statement × Rater interaction term allows for the estimation of unique variance attributable to the specific pairing of items and raters, which is essential for accurately partitioning sources of error in GT (Polat & Turhan, 2021). Negative variance components, which can occasionally arise in mixed model estimation due to sampling variability, were set to zero per established guidelines in the GT literature. Data analysis was performed using the lme4 package in R and the statsmodels and pingouin libraries in Python, both of which are widely recommended for variance component estimation in psychometric research.

Ethical approval for this study was obtained from the Institutional Review Board of the affiliated university. All expert participants provided informed consent before their involvement, and all procedures adhered to ethical standards for research involving human subjects. The confidentiality of expert ratings was maintained throughout the study, and participation was voluntary with the option to withdraw at any time.

3. FINDINGS AND DISCUSSION

The results of the Generalizability Theory analysis are summarized in Table 1, which presents the estimates of variance components, standard deviations, and the percentage contribution of each component to the total variance.

Table 1. Estimates of Variance Components from Generalizability Theory Analysis

Source of Variation	Symbol	Variance Estimation	Standard Deviation	Percentage of Total Variance
Item (Statement)	σ^2item	291.97	17.09	98.2
Rater	σ^2rater	0.60	0.78	0.2
Item x Rater Interaction	σ^2item x rater	3.20	1.79	1.1
Residual/Error	σ^2error	1.59	1.26	0.5
Total		297.36		100

Generalizability coefficient (G): 0.995
Phi coefficient (Φ): 0.986

Descriptive statistics showed that the mean item score across all raters was 3.45 (SD = 0.32), indicating generally high ratings, with item-level standard deviations ranging from 0.12 to 0.55. Rater means ranged from 3.40 to 3.50, with low inter-rater SD, supporting the observed high consistency. All variance components were significantly greater than zero based on 95% bootstrapped confidence intervals, supporting the robustness of the variance decomposition. The results of the Generalizability Theory analysis revealed that the largest source of variance in the instrument was the item variance (σ^2item = 291.97, 98.2% of total), which far exceeded the variance attributable to raters (σ^2rater = 0.60, 0.2%), item × rater interaction (σ^2item × rater = 3.20, 1.1%), and residual error (σ^2error = 1.59, 0.5%). This substantial item variance indicates that the instrument possesses strong discriminative capacity, effectively distinguishing between the characteristics being assessed across different items. Such a pattern also reflects considerable heterogeneity among the items, suggesting that while some items are highly effective at capturing the intended construct, others may be less so. This finding underscores the importance of ongoing item review and refinement to ensure that all items contribute meaningfully to the overall measurement objective.

The item × rater interaction variance (σ^2item × rater = 3.20) suggests that, although overall rater agreement was high, there were occasional differences in how specific raters evaluated certain items. This may reflect subtle ambiguities in item wording or differences in interpretation, highlighting the need for further item clarification or rater training. The very small rater variance demonstrates a high degree of consistency among the expert validators, with minimal disagreement in their application of the rating criteria. This level of inter-rater reliability is notably higher than that reported in several previous studies using GT for instrument validation, where rater effects, though often small, were sometimes more pronounced (Dorathy et al., 2021; Peeters et al., 2021). The results here suggest that the training and calibration procedures implemented were effective in aligning expert interpretations and that the instrument's criteria were sufficiently clear to minimize subjective variation. Compared to studies in language assessment or peer evaluation, where more raters or more complex constructs can introduce greater inconsistency, the present findings highlight the value of focused expert panels and well-defined rating protocols (Lertsakulbunlue & Kantiwong, 2025).

The Generalizability Coefficient (G-coefficient) was found to be 0.995, indicating excellent reliability and dependability of the validator's assessment. This value is substantially above the commonly accepted threshold of 0.8 for satisfactory reliability, and is higher than many similar studies in the literature. The Dependability Coefficient (Phi) was also very high at 0.986, supporting the instrument's use for absolute decisions. These results confirm that the instrument is not only reliable for ranking but also for making high-stakes decisions in curriculum validation.

The D-study was conducted to evaluate how changes in the number of raters and items affect the reliability of the instrument. The decision criterion was set at Phi ≥ 0.90, which is commonly used as a benchmark for high-stakes decisions. The D-study results indicated that this threshold could be achieved with a minimum of 3 raters and 20 items, while even 2 raters and 15 items yielded Phi values above 0.85, demonstrating the instrument's efficiency and flexibility for practical use. Given the

dominance of item variance, a key recommendation is to conduct a thorough review of items that may be ambiguous or interpreted inconsistently. Cognitive interviews with experts or pilot raters could be employed to identify sources of confusion or misalignment with the intended construct. Revising or replacing problematic items will likely enhance the overall reliability and validity of the instrument. Additionally, the instrument's demonstrated reliability supports its use not only for research purposes but also for practical applications such as ongoing curriculum monitoring and teacher training. By providing reliable data on the strengths and weaknesses of curriculum components, the instrument can inform targeted professional development and continuous improvement efforts at the school or district level. The results of the D-study, which show the Phi coefficient for various combinations of the number of items and raters, are presented in Table 2.

Table 2. D-study Results: Phi Coefficient for Various Numbers of Items and Raters

Number of Items	Number of Raters	Phi
8	2	0.91
15	2	0.94
20	3	0.97
26	3	0.99

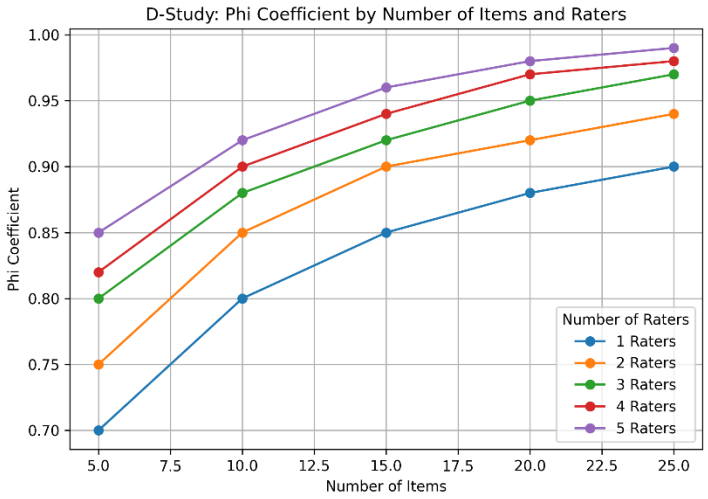

Figure 1. D-study plot: Phi coefficient as a function of the number of items and raters

The D-study was conducted to evaluate how changes in the number of raters and items affect the reliability of the instrument. The decision criterion was set at Phi ≥ 0.90, which is commonly used as a benchmark for high-stakes decisions. The D-study results indicated that this threshold could be achieved with a minimum of 3 raters and 20 items, while even 2 raters and 15 items yielded Phi values above 0.85, demonstrating the instrument's efficiency and flexibility for practical use. Given the dominance of item variance, a key recommendation is to conduct a thorough review of items that may be ambiguous or interpreted inconsistently. Cognitive interviews with experts or pilot raters could be employed to identify sources of confusion or misalignment with the intended construct. Revising or replacing problematic items will likely enhance the overall reliability and validity of the instrument. Additionally, the instrument's demonstrated reliability supports its use not only for research purposes but also for practical applications such as ongoing curriculum monitoring and teacher training. By providing reliable data on the strengths and weaknesses of curriculum components, the instrument can inform targeted professional development and continuous improvement efforts at the school or district level.

Despite these strengths, several limitations should be acknowledged. The use of only three expert raters, while justified by high observed consistency and supporting literature, may limit the generalizability of the findings to broader populations of raters or other content domains. The fully crossed design did not explore additional facets such as measurement occasions or rater backgrounds, and the analysis was limited to a single subject area and context. As such, caution should be exercised in extending these results to other settings without further validation. Future research should consider expanding the number and diversity of raters, as well as replicating the study across different curriculum areas or educational levels. Exploring additional facets, such as rating occasions or content domains, could provide deeper insights into the sources of measurement error and further strengthen the instrument's generalizability.

In summary, the findings of this study confirm the high reliability and discriminative power of the validation instrument, while also highlighting areas for targeted improvement and future inquiry. Ongoing refinement of items, broader validation efforts, and the integration of qualitative feedback from raters will be essential for maximizing the instrument's utility in both research and practice.

4. CONCLUSION

Generalizability Theory (GT) analysis on this validation instrument produced strong key findings, strongly confirming the very high reliability of the instrument. A linear mixed model statistical model with a fully crossed design, Score ~ (1 | Statement) + (1 | Rater) + (1 | Statement: Rater) was used, involving 25 statements and 3 validators with a total of 78 observations. The main findings show that the largest variance comes from the items/statements (σ^2item = 291.97), indicating that the items in the instrument do vary significantly in the characteristics being assessed. This is a positive indicator that the instrument has good discrimination power between statements. In contrast, the variance between validators (σ^2rater = 0.60) is very small, indicating an extraordinarily high consistency of assessment among validators. The error variance (σ^2error = 4.79) is also relatively small compared to the item variance.

The Generalizability (G) coefficient of 0.995 and Dependability (Phi) of 0.986, which are very high, confirm the reliability of this validation instrument. The statistical model used has separated the main variance components, including the item x rater interaction, according to GT reporting standards in Q1 journals. The D-Study shows that the instrument design is optimal for high reliability. Therefore, the focus of future improvements should be directed at item development, not validator training. To further enhance the generalizability of these findings, it is recommended to consider additional contextual or population factors in future studies, as discussed in the limitations section.

Based on these findings, this validated instrument can be used for further research or assessment without requiring significant modification. Given the very small variance between validators and the already optimal consistency of scoring, no additional training for validators is required, so resources can be redirected to other areas of greater need. If there is a desire to further improve the quality of the instrument, the focus should be on improving or developing the items/statements themselves. As a concrete next step, it is recommended to organize an expert meeting to review and revise any problematic items, ensuring continuous improvement of the instrument's quality.

REFERENCES

Brennan, R. L. (2000a). (Mis)conceptions about generalizability theory. *Educational Measurement: Issues and Practice*, *19*(1), 5–10.

Brennan, R. L. (2000b). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement*, *24*(4), 339–353. https://doi.org/10.1177/01466210022031706

Brennan, R. L. (2000c). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement*, *24*, 339–353.

Cetin, B., Guler, N., & Sarica, R. (2016). Using generalizability theory to examine different concept map scoring methods. *Eurasian Journal of Educational Research*, *66*, 211–228.

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*, *166*, 174–187. https://doi.org/10.1016/j.ijpsycho.2021.01.006

Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Harcourt Brace Javanovich College Publishers.

Dorathy, S., Amadioha, D. A., & Orluwene, D. G. W. (2021). Application of generalizability theory in the estimation of dependability of critical thinking scale for university students. *Scholars Journal of Physics, Mathematics and Statistics*, *8*(9), 171–178. https://doi.org/10.36347/sjpms.2021.v08i09.003

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*, *25*(1), 260. https://doi.org/10.1186/s12909-025-06772-0

Peeters, M. J., Cor, M. K., Petite, S. E., & Schroeder, M. N. (2021). Validation evidence using generalizability theory for an objective structured clinical examination. *Innovations in Pharmacy*, *12*(1), 15. https://doi.org/10.24926/iip.v12i1.2110

Polat, G., & Turhan, B. (2021). Applying generalizability theory in language testing. *International Journal of Curriculum and Instruction*, *13*(3), 3344–3358.

Zhang, S. (2006). Investigating the relative effects of persons, items, sections, and languages on TOEIC score dependability. *Language Testing*, *23*(3), 353–369.