UNESA Journal of Chemical Education *ISSN: 2252-9454*
Vol. 14, No. 3, pp. 227–237, September 2025

EVALUATING THE CONSISTENCY OF SENIOR HIGH SCHOOL CHEMISTRY LABORATORY PRACTICE ASSESSMENT USING G-THEORY AND D-STUDY

Yunilia Nur Pratiwi[1,2]* and Raden Rosnawati[1,3]

Research and Evaluation of Education, Postgraduate School, Universitas Negeri Yogyakarta

Department of Chemistry Education, FMIPA, Universitas Negeri Yogyakarta

Department of Mathematics Education, FMIPA, Universitas Negeri Yogyakarta

e-mail: yunilianurpratiwi@uny.ac.id

Abstract

This study aims to evaluate the reliability of chemistry laboratory practice assessments for Grade XII senior high school students by applying Generalizability Theory (G-theory) and Decision Study (D-study). A total of 79 students were involved in the research, each randomly assigned to perform one of six available chemistry laboratory practice tasks: electrolytes and nonelectrolytes, exothermic and endothermic reactions, enthalpy of neutralization between HCl and NaOH, acid-base titration, identification of acidic-basic properties, and electrolysis of $CuSO_4$. Each student was assessed independently by two chemistry teachers based on seven performance criteria: equipment selection, procedure, data reading, analysis, conclusion, cleanliness, and time efficiency. The G-study was conducted using a nested-crossed model in which students were nested within laboratory practice tasks and crossed with raters. The results revealed that variance due to raters (43.2%) and residual error (42.2%) dominated the total score variance, while the variance attributed to students nested within laboratory practice was relatively low (14.6%). The D-study produced a generalizability coefficient ($E\rho^2$) of 0.41 and a dependability index (Φ) of 0.26, indicating low reliability for both relative and absolute decisions. A D-study simulation demonstrated that increasing the number of raters and laboratory practice tasks improved reliability. An optimal configuration of six tasks assessed by nine raters is required to achieve an $E\rho^2 \geq 0.80$. These findings underscore the importance of well-designed assessment systems, consistent rater training, and diverse task coverage to ensure fair and dependable laboratory practice scoring. G-theory and D-study prove to be valuable tools for enhancing the quality of performance-based assessments in science education.
***Keywords**:* practice assessment, reliability, generalizability theory, decision study

INTRODUCTION

Chemistry laboratory assessments in senior high schools are essential for evaluating students' scientific literacy and process skills, including their ability to plan experiments, use laboratory equipment correctly, interpret data, and communicate findings [1]. In Indonesia, such assessments are commonly implemented at the end of Grade XII as part of school-based final examinations. Tasks often include acid-base titrations, calorimetry, electrolysis, and qualitative analysis of household substances [2]. These assessments are generally performance-based and carried out using rubrics that address key aspects such as procedural accuracy, equipment handling, data quality, analytical reasoning, and laboratory discipline [3]. Despite these standardized rubrics, the implementation varies considerably across schools depending on resources, teacher expertise, and institutional policy [2]. Some schools employ multiple raters or structured observation formats, while others rely on a single rater scoring in real time. This variability in implementation raises serious questions about the consistency and reliability of student scores across contexts [4], [5].

In the post-*Ujian Nasional* era, Indonesia's educational system has shifted toward a more decentralized, school-based assessment model. The *Peraturan Menteri Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia No. 21*

year 2022 reinforces this paradigm by emphasizing authentic assessment that aligns with competency-based learning. In this framework, laboratory performance evaluations are expected to serve as valid indicators of student achievement in science [1]. However, concerns have been raised regarding the dependability of such assessments. Variations in scoring outcomes may stem from differences in student ability, inconsistency in rater judgment, or unequal task complexity [3], [4]. These factors can compromise fairness and validity, especially when assessment results are used for high-stakes decisions such as graduation or university admission [6].

Reliability is a cornerstone of high-quality educational assessment, particularly in performance-based tasks such as laboratory examinations. Reliable measurement ensures that scores consistently reflect a student's true level of competence rather than being unduly influenced by external factors such as rater bias, task inconsistency, or environmental variation [7]. In the context of chemistry practical exams, where students are assessed on procedural and analytical skills in real-time, the potential for error introduced by raters and task variations is particularly high. Unlike traditional written tests, performance assessments typically involve greater subjective judgment and require more deliberate efforts to ensure consistent scoring [8], [9]. Without adequate reliability, test scores lose their meaning and may lead to unfair educational decisions.

Classical Test Theory (CTT), which has traditionally been used to evaluate assessment reliability, does not account for multiple sources of measurement error separately [7]. It provides a single error term, conflating various sources such as rater subjectivity, item heterogeneity, and student variability. As a result, CTT lacks diagnostic precision and cannot inform targeted improvements in assessment design. Generalizability Theory (G-Theory), introduced by Cronbach, Gleser, Nanda, and Rajaratnam (1972), overcomes this limitation by allowing researchers to partition observed score variance into multiple facets and interactions, such as persons, raters, tasks, and their combinations [7], [10]. This multifaceted approach makes G-Theory especially well-suited for complex performance assessments like those in chemistry laboratories.

Building upon G-Theory, Decision Studies (D-Studies) enable researchers to explore how changes in assessment design—such as increasing the number of raters or standardizing task procedures—would influence the overall reliability of scores [11]. D-Studies provide a practical framework for balancing trade-offs between cost, feasibility, and measurement precision. This makes G-Theory and D-Study not only diagnostic tools but also valuable in guiding the development of evidence-based assessment policies. In technical and clinical education, G-Theory has been increasingly used to identify dominant sources of variance and improve assessment reliability [11], [12]. For example, Trejo-Mejía et al. (2016) demonstrated that single-rater assessments in medical education frequently failed to meet minimum reliability standards, with generalizability coefficients below 0.70, indicating that multiple-rater systems are necessary to ensure consistent scoring outcomes.

However, despite its growing applications in higher education and medical training contexts, the use of G-Theory in secondary science education—particularly for chemistry laboratory assessments—appears to be underexplored. Research indicates that assessment researchers in education have been generally slow to adopt generalizability theory approaches [14], with applications in school-based performance assessments being particularly rare. This gap is significant given that laboratory work represents a fundamental component of secondary science education globally [15], [1]. Additionally, while G-theory applications exist in controlled assessment environments and large-scale testing programs, few studies have examined its utility in authentic school laboratory settings where multiple sources of measurement error interact in complex ways [16].

To date, no known studies have applied Generalizability Theory (G-Theory) to evaluate the reliability of chemistry laboratory assessments in Indonesian high schools. This gap is critical, given

the increasing reliance on school-based assessments for determining student progression and graduation [17]. Without empirical evidence on the dependability of these assessments, educational stakeholders may be making consequential decisions based on potentially unreliable data [6], [4]. Furthermore, teachers may lack the necessary feedback to refine their assessment practices and ensure that student performance is judged fairly and consistently [7], [3].

In response to this gap, the present study aims to evaluate the consistency of chemistry laboratory performance assessments conducted in Grade XII of Indonesian senior high schools by applying Generalizability Theory and Decision Study. Specifically, the study seeks to identify and quantify sources of score variance related to students, raters, and tasks, and to simulate how changes in the number of raters or laboratory practices might affect overall reliability. By addressing both methodological and practical concerns, this research contributes to the development of more dependable, equitable, and data-driven assessment systems in science education.

METHOD

This study employed a quantitative approach with a non-experimental descriptive design to apply Generalizability Theory (G-theory) and Decision Study (D-study) within the context of chemistry laboratory practice assessment. The primary objective was to evaluate the reliability of the assessment system by examining various sources of variance, particularly those related to students, laboratory practice tasks, and raters.

The research was conducted at a senior high school in Malang, East Java during the second semester of the 2023/2024 academic year. The participants consisted of 79 Grade XII students enrolled in the science (MIPA) program, who were required to take part in a chemistry laboratory practice examination as one of the graduation requirements. The laboratory practice activities were carried out in the school's laboratory according to a predetermined schedule.

Each student was required to randomly select one laboratory practice task from a set of available options. This random assignment was intended to simulate the diversity of examination conditions and to evaluate the reliability of assessment using a nested design, in which students were nested within laboratory practice tasks. During the assessment process, each student was evaluated independently by two chemistry teachers who were unaware of each other's ratings. Both teachers served as crossed raters, meaning they assessed all students regardless of the laboratory practice type.

The assessment instrument employed was a performance-based rubric that has been previously undergone content validation by a team of science subject teachers. This instrument was currently employed to evaluate practical exams in Chemistry, Physics, and Biology at the end of senior high school level. At present, the observation sheets and scoring rubrics are still limited-use, primarily within the internal context of the school. The rubric comprised seven assessment criteria, including equipment selection, procedural execution, data interpretation, analysis or discussion, conclusion formulation, laboratory cleanliness, and time management. Each criterion was rated on a 5-point Likert scale, and the total score was subsequently converted into a standardized scale with a maximum score of 100.

The score data obtained from two raters for each student were processed and analyzed using the Generalizability Theory approach, employing a (student nested within laboratory practice) × rater design. The data were structured in a long format and analyzed using the R statistical software, utilizing the gtheory package. The G-study was conducted to estimate the variance components attributable to the student–laboratory practice nested facet, rater, and residual error. Subsequently, the D-study was carried out to compute the generalizability coefficient ($E\rho^2$) and the index of dependability (Φ), and to evaluate hypothetical scenarios involving the addition of raters and laboratory practice tasks to examine their effects on improving assessment reliability.

Table 1. Guidelines for Interpreting G-Theory Coefficients

Coefficient	Typical Value Range	Interpretation
$E\rho^2$ (Generalizability Coefficient)	≥ 0.80	Very high reliability; suitable for high-stakes decision-making and comparing student performance
	0.70 – 0.79	Acceptable reliability; suitable for formative evaluation
	0.60 – 0.69	Moderate reliability; use with caution
	< 0.60	Low reliability; not recommended without improvement
Φ (Index of Dependability)	≥ 0.80	High reliability; suitable for absolute decisions (e.g., pass/fail)
	< 0.70	Insufficient for dependable absolute decisions

Note: These thresholds are commonly referenced in G-theory literature, including [18], [19], [20], and may vary slightly depending on context, stakes, and intended use of the assessment scores.

The interpretation of the reliability coefficients **$E\rho^2$** and **Φ** in this study refers to the threshold values recommended in the Generalizability Theory literature, namely **≥ 0.70** for moderate reliability and **≥ 0.80** for high reliability, particularly in the context of high-stakes decision-making [18], [19], [20].

RESULTS AND DISCUSSION

A total of 79 twelveth-grade senior high school students participated as research subjects in this study by taking part in a chemistry laboratory practice examination, which served as a component of their end-of-semester assessment. Each student was asked to select one laboratory practice topic through a closed random draw system. Six types of laboratory practice tasks were available, namely: (1) Electrolyte and Nonelectrolyte Solutions; (2) exothermic and endothermic reactions; (3) determination of enthalpy change in the neutralization reaction between HCl and NaOh; (4) acid-base titration; (5) identification of acid-base properties of solutions; and (6) electrolysis of $CuSO_4$ solution. This random selection was designed to simulate the diversity of laboratory practice contexts that might influence assessment outcomes and to examine the stability of student scores under varying task conditions.

After completing their assigned laboratory practice, each student was directly assessed by two chemistry teachers who served as raters. Both raters evaluated the students independently and without consultation, in order to maintain assessment objectivity and minimize potential collaborative bias.

The assessment was conducted using a performance-based rubric comprising seven aspects of practical skills: equipment selection, procedural execution, data interpretation, analysis or discussion of results, conclusion drawing, laboratory cleanliness, and time management. Each aspect was rated on a 1–5 scale by each rater, and the total score was calculated by summing all aspects. The final scores from each rater were then used to analyze the reliability of the assessment using Generalizability Theory, with the support of the R statistical software.

The G-study was conducted to evaluate the sources of variance in the chemistry laboratory practice assessment using a design model in which students were nested within laboratory practice tasks and assessed by raters (teachers) in a crossed manner. Based on the analysis, it was found that the total score variance assigned by two chemistry teachers to 79 Grade XII students, each of whom completed one of six laboratory practice tasks, consisted of three main components: variance among students nested within laboratory practice tasks, variance attributable to raters, and residual variance. The following presents the results of the G-Theory analysis.

```
  source          var       percent n
1 SISWA.PRAKTIKUM  6.283511    14.6 1
2           RATER 18.538185    43.2 1
3        Residual 18.120091    42.2 1
```

```
d_model$generalizability
```

```
# Koefisien Eρ²
[1] 0.4095212

d_model$dependability
# Koefisien Φ
[1] 0.255296
```

The generalizability theory analysis of chemistry laboratory assessment scores revealed a troubling variance decomposition pattern that fundamentally challenges the psychometric integrity of the current evaluation system. The analysis demonstrated that measurement error dominated true score variance, with rater effects contributing 43.2% of total variance, residual variance accounting for 42.2%, and student-by-practical interaction representing only 14.6% of the observed score variation. This distribution indicates that student scores reflect measurement artifacts more than actual chemistry competencies, rendering the assessment system psychometrically inadequate for educational decision-making purposes. The magnitude of error variance observed substantially exceeds acceptable thresholds established in the measurement literature, suggesting systematic flaws in assessment design and implementation that require immediate attention.

The dominance of rater variance (σ^2r = 18.54, 43.2%) represents the most critical measurement flaw identified in this study, indicating severe inconsistencies in scoring practices across different evaluators. This finding extends beyond typical inter-rater reliability concerns documented in performance assessment literature, approaching levels that would render score interpretations indefensible for consequential educational decisions. The substantial rater effects observed suggest fundamental problems in rubric interpretation, scoring criteria application, or systematic differences in rater severity patterns that persist despite existing training protocols. Compared to well-designed performance assessments where rater variance typically contributes 15-25% of total variance [22], the current results indicate that student scores are more reflective of "who scored them" rather than their demonstrated chemistry proficiency levels. Such pronounced rater dependency threatens the fairness and validity of assessment outcomes, particularly in high-stakes contexts where scores influence academic progression and university admission decisions.

The substantial residual variance (42.2%) represents an equally concerning finding, encompassing unidentified and uncontrolled error sources that were not explicitly modeled in the generalizability study design. This elevated residual component, which exceeds the 20-30% typically observed in well-controlled assessment contexts [19], suggests the presence of multiple confounding factors affecting score reliability. These unmodeled sources likely include temporal effects from varying testing conditions, task-specific complexities not captured by the student-by-practical interaction, equipment and resource variability across different laboratory settings, and administrative inconsistencies in testing procedures. The magnitude of residual variance may also indicate model misspecification, suggesting that additional facets such as task difficulty, laboratory conditions, or school-level effects should have been explicitly incorporated into the analysis design [23]. This finding underscores the complexity of laboratory assessment contexts and the challenge of achieving measurement consistency across diverse educational environments.

The obtained reliability coefficients provide compelling evidence of the assessment system's inadequacy for both relative and absolute decision-making contexts. The generalizability coefficient of G = 0.41 indicates that only 41% of observed score variance reflects true differences in student chemistry competencies, falling substantially below the 0.70-0.80 threshold typically required for educational assessments [24]. This level of reliability renders the assessments unsuitable for rank-ordering students or making comparative evaluations that influence academic opportunities. Even more concerning is the dependability coefficient of D = 0.26, which indicates that 74% of score variance consists of measurement error when making absolute judgments about student proficiency levels. Such

poor dependability makes the assessments fundamentally unreliable for mastery decisions, certification purposes, or minimum competency determinations, as pass/fail classifications would be largely arbitrary and potentially discriminatory [18].

These findings contribute to the growing body of evidence questioning the psychometric adequacy of performance-based assessments in science education, while revealing more severe measurement problems than typically reported in international literature. Previous research by Ruiz-Primo and Shavelson (1996) documented balanced variance distributions in hands-on science assessments, with student variance contributing 40-50% of total variance in well-designed systems. The inversion of this pattern observed in the current study—where error sources dominate over true score variance—suggests that Indonesian chemistry laboratory assessments may require fundamental redesign rather than incremental improvements. This finding aligns with concerns raised by Gott and Duggan (1995) and Hofstein and Lunetta (2004) regarding the challenges of implementing reliable practical assessments, but demonstrates that these challenges may be more severe in contexts with limited assessment infrastructure and training resources. The substantial deviation from international benchmarks indicates that local contextual factors, including teacher preparation, resource availability, and institutional support systems, play critical roles in assessment quality that deserve greater research attention.

The psychometric evidence presented necessitates immediate reforms in assessment practice and educational policy to protect student interests and maintain system credibility. Given the unacceptable reliability levels, scores from current chemistry laboratory assessments should not be used for high-stakes decisions affecting student academic futures until substantial improvements are implemented. The dominant rater effects demand comprehensive training programs focusing on rubric interpretation, scoring calibration, and bias recognition, supported by ongoing quality assurance monitoring. Implementation of independent multiple-rater scoring protocols with systematic consensus-building procedures represents a critical short-term intervention to reduce rater-specific variance. However, the high residual variance suggests that more fundamental revisions of task design, scoring procedures, and administration protocols may be necessary to achieve acceptable measurement quality. These reforms require coordinated efforts among curriculum developers, teacher educators, school administrators, and policy makers to ensure sustainable improvements in assessment practice.

These findings reinforce the concern that conducting a single-session laboratory practice assessment involving two uncalibrated raters carries a high risk of scoring inconsistency. The substantial variance attributed to raters reflects differing interpretations of the rubric or scoring criteria among teachers, despite prior validation of the instrument. This underscores the importance of structured rater training to establish shared understandings and improve inter-rater consistency, as strongly advocated in the literature on performance assessment in science education [27].

In addition, the substantial residual variance indicates that non-systematic factors during laboratory practice implementation—such as laboratory conditions, scheduling, or student anxiety—also contributed to score instability. Therefore, the assessment design should ideally incorporate more than one laboratory practice session or task type, along with a rater pool exceeding two assessors, to enhance reliability and reduce error variance. A follow-up Decision Study can be conducted to simulate the effect of increasing the number of raters while maintaining six laboratory practice tasks, in order to determine the optimal assessment configuration within the constraints of available resources.

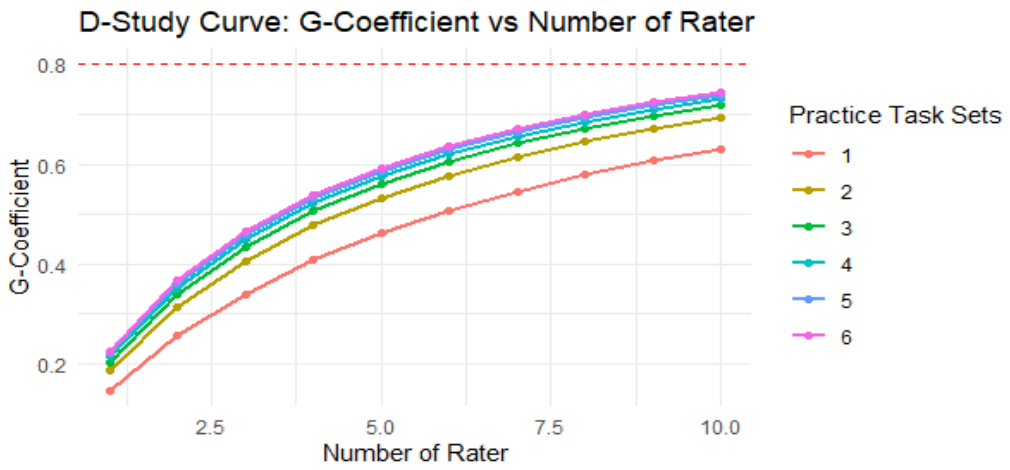

Figure 1. Simulation of D-Study

The decision study results depicted in the G-coefficient curves provide critical insights for optimizing chemistry laboratory assessment design and reveal the complex interplay between the number of raters and practice task sets in achieving acceptable measurement reliability. The horizontal threshold line at G = 0.80 represents the conventional minimum standard for educational assessments involving consequential decisions, as established by Nunnally and Bernstein (1994) and reinforced by contemporary measurement standards. The systematic examination of these optimization curves demonstrates that achieving acceptable reliability in chemistry laboratory assessments requires careful consideration of both human and material resources, with profound implications for assessment feasibility and educational policy.

The most striking finding from the decision study analysis is the dramatic impact of practice task diversity on measurement reliability, as evidenced by the substantial separation between the curve representing a single practice task set (red line) and those representing multiple task sets. Even with the maximum number of raters examined (n=10), the single-task condition achieves only G = 0.63, falling substantially below acceptable thresholds and confirming that rater multiplication alone cannot compensate for insufficient task sampling. This finding aligns with seminal work by Shavelson and Webb (1991), who demonstrated that task diversity often contributes more to measurement reliability than rater consistency in performance-based assessments. However, the magnitude of this effect in chemistry laboratory contexts exceeds that reported in previous studies, where single-task designs typically achieve G-coefficients in the 0.70-0.75 range with adequate rater numbers [18]. The particularly poor performance of single-task designs in our context suggests that chemistry laboratory competencies may be more task-specific than previously assumed, requiring broader sampling across different experimental procedures to achieve generalizable competency inferences.

The convergence of curves representing three or more practice task sets reveals a critical threshold effect that has significant implications for assessment design optimization. Beyond three practice tasks, additional task sets provide minimal improvement in measurement reliability, regardless of rater numbers employed. This finding contradicts the linear relationship between task sampling and reliability suggested by classical generalizability theory formulations, instead supporting a saturation model where measurement benefits plateau after achieving adequate content sampling. The convergence pattern observed is consistent with recent findings by Vispoel et al. (2018) in performance assessment contexts, who reported similar diminishing returns in multi-task designs. However, the rapid convergence observed in chemistry laboratory assessments occurs earlier than the five-to-seven task threshold typically reported in other domains, suggesting that chemistry practical competencies may be more efficiently sampled than competencies in fields requiring more diverse skill demonstrations.

The optimization curves demonstrate clear diminishing returns in reliability improvement as rater numbers increase beyond four evaluators, with particularly steep improvements occurring between one and three raters before leveling substantially. This pattern confirms predictions from generalizability theory regarding the relative efficiency of rater sampling compared to other design facets, but the specific shape of these curves provides practical guidance rarely available in the literature. With three practice task sets, acceptable reliability ($G \geq 0.80$) is achieved with approximately four raters, while designs employing four or more task sets can reach this threshold with three raters. These findings closely parallel results reported by Lane and Stone (2006) in large-scale performance assessments, where optimal designs typically employed three to four trained raters for complex performance tasks. The similarity of optimal rater numbers across different assessment contexts suggests robust principles of rater sampling that transcend specific content domains.

The economic implications of these optimization results are profound and extend beyond simple reliability considerations to encompass practical feasibility and sustainability of assessment programs. Employing four raters across three practice task sets requires twelve independent scoring events per student, representing a substantial resource commitment that may be prohibitive for routine assessment implementation. The marginal improvement in reliability achieved by moving from three to four raters (approximately 0.02-0.03 G-coefficient units) must be weighed against the 33% increase in scoring costs and logistical complexity. This cost-benefit analysis aligns with recommendations by Cardinet et al. (2010), who emphasized that optimization decisions should balance psychometric quality with practical constraints. However, the current analysis extends beyond traditional optimization by demonstrating that strategic task selection may offer more cost-effective reliability improvements than rater multiplication, suggesting that curriculum design and assessment planning should prioritize diverse practical experiences over intensive scoring protocols.

The interaction effects revealed in these curves have important implications for assessment validity that extend beyond reliability considerations alone. The poor performance of single-task designs threatens not only score consistency but also the representativeness of competency inferences, as students may demonstrate highly variable performance across different laboratory procedures despite consistent scoring. This validity threat is particularly concerning in chemistry education, where laboratory competencies encompass diverse skills ranging from quantitative analysis to synthetic procedures, each requiring different cognitive and psychomotor abilities. The finding that three practice tasks appear sufficient for reliable measurement suggests that carefully selected laboratory experiences can capture the essential variance in student competencies without overwhelming assessment systems. This conclusion supports contemporary trends toward competency-based assessment design, where strategic sampling of key skills takes precedence over comprehensive coverage of all possible laboratory procedures.

The implications of these optimization results for assessment policy and practice are immediate and consequential, particularly given the current state of reliability inadequacy revealed in the initial generalizability study. Implementation of optimal assessment designs (three to four raters, three practice task sets) would require fundamental restructuring of current assessment practices, including coordinated training programs, standardized scoring protocols, and institutional resource allocation. The decision study results suggest that such investments would yield substantial improvements in measurement quality, potentially transforming currently unreliable assessments into psychometrically defensible evaluation systems. However, successful implementation would require sustained institutional commitment and coordination across multiple stakeholder groups, challenges that have historically impeded assessment reform initiatives in educational contexts.

Several methodological limitations constrain the interpretation and generalizability of these optimization results, pointing toward important directions for future research and development. The decision study curves assume that variance component estimates remain stable across different design configurations, an assumption that may not hold if rater training effectiveness varies with group size or if task difficulty interactions emerge with different sampling strategies. Additionally, the analysis focuses exclusively on generalizability coefficients without examining dependability coefficients for absolute decision-making contexts, potentially overlooking important optimization differences for criterion-referenced interpretations. Future research should investigate whether these optimization patterns hold across different chemistry content areas, educational levels, and cultural contexts, as the current results derive from a specific institutional and curricular context that may not represent broader chemistry education practices.

The decision study analysis provides clear guidance for immediate assessment improvement while highlighting the complex resource allocation decisions facing chemistry educators seeking to implement reliable performance assessments. The finding that optimal designs require significantly more resources than current practices underscores the fundamental tension between measurement quality and practical feasibility that pervades educational assessment. Nevertheless, the systematic nature of these optimization results offers evidence-based guidance for assessment designers willing to invest in measurement quality, demonstrating that substantial reliability improvements are achievable through strategic design modifications. The convergence of reliability curves around the 0.80 threshold provides a concrete target for assessment development efforts, while the clear identification of optimal design parameters eliminates much of the guesswork traditionally associated with performance assessment planning.

CONCLUSION

This generalizability theory analysis reveals fundamental psychometric inadequacies in current chemistry laboratory assessment practices, with measurement error dominating true score variance (rater effects 43.2%, residual variance 42.2%). The obtained generalizability coefficient of 0.41 and dependability coefficient of 0.26 fall substantially below acceptable thresholds, rendering current assessments unsuitable for high-stakes educational decisions affecting student academic progression.

The decision study optimization demonstrates that achieving acceptable reliability ($G \geq 0.80$) requires substantial resource investments—three to four trained raters across three practice task sets—representing a fundamental shift from current single-rater approaches. The dramatic performance differences between single-task and multi-task designs underscore the critical importance of content sampling in laboratory competency assessment, challenging assumptions about the generalizability of chemistry practical skills across experimental contexts.

These findings contribute to international discourse on performance-based assessment reliability while highlighting unique challenges in developing educational contexts, where reliability problems exceed those in well-resourced systems. The successful application of generalizability theory methodology demonstrates the value of sophisticated psychometric approaches for diagnosing assessment quality issues across diverse educational environments.

The implications for practice require coordinated action: immediate implementation of multi-rater protocols and comprehensive training programs to address severe rater effects, coupled with longer-term reforms addressing assessment design, task selection, and quality monitoring systems. The substantial resource requirements for optimal designs necessitate careful consideration of sustainability and institutional capacity.

Several limitations constrain generalizability of findings, including the single-institution design and exclusive focus on reliability without validity evidence. Future research should employ multi-site longitudinal designs, investigate innovative scoring approaches to reduce rater dependency, and integrate cognitive diagnostic models with generalizability theory for enhanced competency profiling.

In conclusion, this study demonstrates that rigorous psychometric evaluation can reveal serious assessment quality problems while providing evidence-based improvement guidance.

The findings underscore the critical importance of investing in assessment quality as fundamental to educational effectiveness, highlighting complex resource decisions facing institutions implementing reliable performance-based evaluations. The chemistry education community must prioritize systematic assessment improvement to protect student interests and maintain educational evaluation credibility.

REFERENCES

1. Hofstein, A., and Lunetta, V. N. 2004. The Laboratory in Science Education: Foundations for The Twenty-First Century. *Science Education*, Vol. 88, No. 1, pp. 28–54.
2. Kang, N. H., and Wallace, C. S. 2005. Secondary Science Teachers' Use of Laboratory Activities: Linking Epistemological Beliefs, Goals, and Practices. *Science Education*, Vol. 89, No. 1, pp. 140–165.
3. Harlen, W., and Qualter, A. 2018. *The Teaching Of Science in Primary Schools (7th ed.)*. Routledge.
4. Gipps, C. V. 1994. *Beyond Testing: Towards a Theory of Educational Assessment*. Routledge.
5. Abrahams, I., and Millar, R. 2008. Does Practical Work Really Work? A Study of The Effectiveness of Practical Work as A Teaching and Learning Method in School Science. *International Journal of Science Education*, Vol. 30, No. 14, pp. 1945–1969.
6. Brookhart, S. M. 2013. *How to Create and Use Rubrics for Formative Assessment and Grading*. ASCD.
7. Fan, X., and Sun, S. 2014. Generalizability Theory as a Unifying Framework of Measurement Reliability in Adolescent Research. *The Journal of Early Adolescence*, Vol. 34, No. 1, pp. 38–65.
8. Darling-Hammond, L., and Adamson, F. 2013. *Developing Assessments of Deeper Learning: The Costs and Benefits of Using Tests That Help Students Learn*. Stanford Center for Opportunity Policy in Education.
9. Palm, T. 2008. Performance Assessment and Authentic Assessment: A Conceptual Analysis of The Literature. *Practical Assessment, Research & Evaluation*, Vol. 13, No. 4, pp. 1–11.
10. Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratnam, N. 1972. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. John Wiley & Sons.
11. Monteiro, S., Sullivan, G. M., and Chan, T. M. 2020. Generalizability Theory Made Simple(R): An Introductory Primer to G-Studies. *Journal of Graduate Medical Education*, Vol. 12, No. 4, pp. 365–370.
12. Bloch, R., and Norman, G. 2012. Generalizability Theory for The Perplexed: A Practical Introduction and Guide: AMEE Guide No. 68. *Medical Teacher*, Vol. 34, No. 11, pp. 960–992.
13. Trejo-Mejía, J. A., Sánchez-Mendiola, M., Méndez-Ramírez, I., and Martínez-González, A. 2016. Reliability Analysis of The Objective Structured Clinical Examination Using Generalizability Theory. *Medical Teacher*, Vol. 38, No. 1, pp. 44–49.
14. Briesch, A. M., Swaminathan, H., Welsh, M., and Chafouleas, S. M. 2014. Generalizability Theory: A Practical Guide to Study Design, Implementation, and Interpretation. *Journal of School Psychology*, Vol. 52, No. 1, pp. 13–35.
15. National Research Council. 2006. *America's Lab Report: Investigations in High School Science*. The National Academies Press.
16. Agustian, H. Y., and Seery, M. K. 2017. Reasserting The Role of Pre-Laboratory Activities in Chemistry Education: A Proposed Framework for Their Design. *Chemistry Education Research and Practice*, Vol. 18, No. 4, pp. 518–532.
17. Kemendikbudristek. 2022. *Peraturan Menteri Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia Nomor 21 Tahun 2022 tentang Standar Penilaian Pendidikan*. Jakarta: Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi.
18. Brennan, R. L. 2001. *Generalizability Theory*. Springer-Verlag.

19. Shavelson, R. J., and Webb, N. M. 1991. *Generalizability Theory: A Primer*. Sage Publications.

20. American Educational Research Association. 2014. *American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

21. Kane, M. T. 2006. *Validation.* In R. L. Brennan (Ed.), Educational Measurement (4th ed., pp. 17–64). American Council on Education/Praeger.

22. Lane, S., and Stone, C. A. 2006. *Performance Assessment*. In R. L. Brennan (Ed.), Educational Measurement (4th ed., pp. 387-431). American Council on Education/Praeger.

23. Cardinet, J., Johnson, S., and Pini, G. 2010. *Applying Generalizability Theory Using EduG*. Routledge.

24. Nunnally, J. C., and Bernstein, I. H. 1994. *Psychometric Theory* (3rd ed.). McGraw-Hill.

25. Ruiz-Primo, M. A., and Shavelson, R. J. 1996. Problems and Issues in The Use of Concept Maps in Science Assessment. *Journal of Research in Science Teaching*, Vol. 33, No. 6, pp. 569–600.

26. Gott, R., and Duggan, S. 1995. *Investigative Work in The Science Curriculum: Developing Science and Technology Education*. Open University Press.

27. Vispoel, W. P., Morris, C. A., and Kilinc, M. 2018. Applications of Generalizability Theory and Their Relations to Classical Test Theory and Structural Equation Modeling. *Psychological Methods*, Vol. 23, No. 1, pp. 1–26.