EDUTECH

Jurnal Teknologi Pendidikan

Journal homepage https://ejournal.upi.edu/index.php/edutech

Analyzing Variance Sources in Indonesian Microteaching Assessment through Generalizability Theory

Nuridayanti, Yetti Supriyati, Ilham Falani, dan Andi Muhammad Ishak
Universitas Negeri Makassar, Indonesia
E-mail: nuridayanti@unm.ac.id

ABSTRACT

Objective This study aims to examine the reliability of student performance assessments in microteaching using G-Theory, and to identify the primary sources of score variance that may affect the accuracy of evaluation outcomes.

Methods A fully crossed p × r × i design was employed, involving 32 students, 2 raters, and 124 observation items representing ten core teaching skills. Secondary data were analyzed using EduG 6.1e software. The analysis included both a G-Study, to estimate variance components, and a D-Study, to simulate changes in reliability based on different rater configurations.

Results The results of the G-Study revealed that person variance (24.9%) and person × rater interaction (23.5%) were the dominant sources of score variability, while item and rater variances were negligible. The residual three-way interaction (person × item × rater) accounted for 50.9% of the total variance, indicating substantial unexplained error. The initial generalizability coefficient was 0.68, categorized as moderate reliability. The D-Study demonstrated that increasing the number of raters to three or more significantly improved reliability, with the coefficient reaching 0.76-0.82, and reduced the standard error of measurement.

Conclusion The study concludes that microteaching assessments are susceptible to reliability issues due to rater inconsistency. Applying G-Theory provides empirical justification for improving rater calibration and increasing the number of rater.

ARTICLE INFO

Article History:
Submitted/Received 5 Juni 2025
First Revised 19 Juni 2025
Accepted 28 Juni 2025
First Available online 01 Okt 2025
Publication Date 01 Okt 2025

Keyword:
Decision Study,
Generalizability Theory,
Performance Assessment,
Microteaching,
Rater Reliability

1. INTRODUCTION

The education of prospective teachers demands a learning system that is not only theoretical but also practice-oriented to ensure their readiness in facing real classroom situations. In this context, microteaching has become one of the most widely implemented training approaches, as it provides direct teaching experience on a limited scale. Microteaching enables students to develop and demonstrate pedagogical skills in a structured and controlled setting, serving as an essential means of building foundational professional competencies.

As a training method, microteaching has been shown to offer numerous benefits in enhancing the teaching quality of pre-service teachers. Various studies have indicated that microteaching improves technical teaching skills, such as classroom management, instructional delivery techniques, and student interaction (Altammar & Aljassar, 2021; Mishra, 2024; Özcan & Gerçek, 2019). Moreover, it contributes to building students' self-confidence by providing opportunities for self-reflection and allowing them to receive feedback from peers and supervising lecturers (Deshpande & Shastri, 2020). The practice of microteaching has also proven effective in reducing anxiety commonly experienced by students when confronted with actual teaching situations (Şen, 2009). Additionally, microteaching deepens understanding of instructional content, particularly in complex subject areas such as mathematics and physics (Komolafe et al., 2020; Mishra, 2024).

The effectiveness of microteaching in enhancing pre-service teachers' competencies is inherently linked to how the practice is evaluated. One of the primary requirements for a successful evaluation process is the availability of valid and reliable instruments. The use of standardized tools such as the Microteaching Assessment Questionnaire (MAQ) facilitates a more systematic and objective assessment (González-Mélendez et al., 2023; Padmadewi & Artini, 2019). Furthermore, feedback and reflection play a crucial role in reinforcing the learning process. Students who receive meaningful feedback from lecturers or peers, and engage in reflective practices regarding their performance, tend to demonstrate more significant improvements in their teaching strategies (Aruğaslan, 2025; Sudrajat et al., 2024).

Despite the numerous benefits outlined, microteaching practices are not without limitations. Several studies have highlighted that interaction skills and the ability to summarize content remain common areas of weakness in student performance (Deshpande & Shastri, 2020). Moreover, inconsistencies between the skills exhibited in online and offline learning environments reveal the need for an assessment approach capable of addressing such situational variations (Raharjo et al., 2025). This underscores the importance of having an evaluation system that is not only accurate but also adaptable to diverse implementation contexts.

In response to these various challenges, several recommendations have been proposed by researchers to enhance the effectiveness of microteaching. Pre-service teachers are encouraged to engage more frequently in microteaching sessions and receive intensive supervision to ensure consistent improvement in skill mastery (Mishra, 2024). Strengthening the integration between pedagogical theory and hands-on practice is also essential, as it has been shown to significantly improve overall teaching quality (Komolafe et al., 2020). Additionally, assessments should be conducted repeatedly and continuously to ensure that the development of teaching skills can be monitored more accurately and systematically (Koech & Mwei, 2019).

Nevertheless, the success of microteaching depends not only on its implementation but also on how the assessment process is conducted. Microteaching assessments are

DOI: https://doi.org/10.17509/e.v24i3.86980
P-ISSN 0852-1190 E-ISSN 2502-0781

often subjective, as they rely on direct observation of student teaching behavior by evaluators. Assessment bias may occur due to evaluator characteristics, such as a tendency to be overly lenient or overly strict. In their study, Jones & Bergin (2019) found that approximately 12% of evaluators exhibited severe rating bias. When assessments involve multiple raters and a large number of instrument items, score variability may not solely reflect students' true teaching abilities, but may also be influenced by rater perceptions and item characteristics. Practitioners must be able to accurately identify the factors involved in their assessment applications and appropriately classify them either as objects of study or as sources of measurement error (Cardinet et al., 2010).

Generalizability Theory (G-Theory)

Generalizability Theory (G-Theory) is a statistical framework used to assess the reliability and consistency of measurement scores across various conditions and facets of a measurement procedure. It is an extension of *Classical Test Theory* (CTT) that allows for the separation of multiple sources of error variance, thereby providing a more comprehensive understanding of measurement reliability (Brennan, 2009; Gudiato et al., 2024; Hendrickson & Yin, 2018).

One of the core concepts in G-Theory is the distinction between dependability and generalizability. Dependability refers to the consistency of scores across conditions, while generalizability pertains to the extent to which findings can be generalized to a broader universe of observations (Briesch et al., 2014; Matt & Sklar, 2015). Unlike CTT, which typically considers only a single source of error, G-Theory accommodates multiple sources of error simultaneously, such as raters, items, and occasions (Brennan, 2009; Teker et al., 2015). To identify and estimate these sources of measurement error, G-Theory employs analysis of variance (ANOVA) techniques to estimate variance components associated with different facets of the measurement process, such as the individuals assessed, the items, and the raters (Brennan, 2001, 2010; Teker et al., 2015).

In practice, G-Theory is widely applied in educational measurement to enhance the design and reliability of assessments. It is valuable for both relative decision-making (e.g., comparing individuals) and absolute decision-making (e.g., determining qualification or competency) (Briesch et al., 2014, 2016; Hendrickson & Yin, 2018). Compared to CTT, G-Theory offers several advantages, including the ability to simultaneously estimate multiple sources of error, resulting in more detailed and accurate assessments of measurement reliability (Brennan, 2009, 2010; Teker et al., 2015). Moreover, G-Theory is flexible, as it does not require parallel test forms and can accommodate unbalanced measurement designs, making it particularly suitable for complex measurement scenarios (Clayson & Miller, 2017). G-Theory also provides an integrated approach that bridges the concepts of reliability and validity, thereby enabling more trustworthy measurement designs (Brennan, 2001; Matt & Sklar, 2015).

Brennan (2001) stated that G-Theory is an extension of CTT that allows for a more flexible and multifaceted approach to reliability analysis. In G-Theory, measurement scores are viewed as the result of a combination of various sources of variance, rather than solely random error as assumed in CTT. Brennan emphasized that every measurement occurs within a specific context involving one or more facets, such as raters, items, time, and administrative conditions. Consequently, score reliability is no longer treated as a fixed value but rather as a function of the measurement design employed. Through this approach, G-Theory not only provides more realistic reliability estimates but also enables researchers to design follow-up studies known as decision

DOI: https://doi.org/10.17509/e.v24i3.86980
P-ISSN 0852-1190 E-ISSN 2502-0781

studies to evaluate and optimize the assessment structure for more efficient and accurate measurement-based decision making.

In conclusion, G-Theory offers a robust and flexible framework for assessing measurement reliability and validity across various domains. By accounting for multiple sources of error and offering a comprehensive approach to measurement design, G-Theory enhances the dependability and generalizability of research findings (Brennan, 2010; Briesch et al., 2016; Hendrickson & Yin, 2018; Matt & Sklar, 2015; Teker et al., 2015).

From a methodological perspective, the assessment of microteaching practices has traditionally relied on Classical Test Theory (CTT), which presents fundamental limitations due to its assumption that observed scores are simply the result of true ability and a single undifferentiated error term. CTT is unable to distinguish between sources of variance stemming from raters, items, or the interactions among these components. Therefore, Generalizability Theory (G-Theory) was developed as a more flexible and comprehensive alternative. G-Theory allows for variance component analysis to identify and estimate the extent to which each facet contributes to the total score. G-Theory consists of two key stages: the Generalizability Study (G-study), which aims to estimate the magnitude of different sources of measurement error, and the Decision Study (D-study), which uses the information from the G-study to optimize the design of measurement procedures for better reliability and efficiency in decision-making (Brennan, 2001).

Monteiro et al. (2019) describe a G study as a type of study aimed at identifying the variance components that influence assessment scores referred to as facets such as the contributions of participants, raters, and items to the final score. This study enables a more comprehensive estimation of reliability compared to classical approaches, as it takes into account various sources of score instability. Within the G-Theory framework, standard reliability questions such as interrater reliability, test-retest reliability, or a combination of both are reinterpreted in terms of the extent to which scores can be generalized across raters, occasions, or both simultaneously.

In G-Theory, facets refer to the dimensions of measurement (e.g., participants, raters, items, or other components that may serve as sources of error), and can be classified as fixed or random depending on the assessment design (Dzakadzie & Quansah, 2023). Understanding the nature of facets is essential for determining whether the findings of a study can be generalized to other contexts. The G-coefficient is used to estimate the extent to which scores can be generalized, either relatively (within a specific context) or absolutely (to a broader context). In practice, the use of the absolute coefficient is often recommended as a more conservative approach, particularly in formative and summative decision-making processes (Monteiro et al., 2019).

Despite its advantages, the application of G-Theory in assessment especially in microteaching evaluation remains limited. The evaluation of pre-service teachers' teaching practice often relies on only two or three raters and uses instruments with numerous items, without a thorough analysis of the sources of variance. When facets such as raters and items are not properly controlled or analyzed, the assessment becomes prone to bias. For example, high variability among raters can compromise score fairness, while inconsistent items may lead to evaluation outcomes that do not accurately represent students' actual teaching competence.

Therefore, it is essential to conduct an analysis using G-Theory to identify and precisely estimate the various sources of score inconsistency in microteaching assessment such as variance originating from participants, raters, and assessment

DOI: https://doi.org/10.17509/e.v24i3.86980
P-ISSN 0852-1190 E-ISSN 2502-0781

items, as well as their interactions. This approach enables assessment designers to gain a more comprehensive understanding of the reliability of the evaluation system and provides a strong empirical foundation for improving the quality of pre-service teacher assessment. In Indonesia, where microteaching assessments often rely on subjective observation without systematic rater calibration, the application of G-Theory is particularly relevant to ensure fairness, accuracy, and accountability in academic decision-making concerning students.

Research Gap and Study Novelty

This research plays a strategic role in the development of performance-based teaching practice evaluation in Indonesian higher education, particularly in microteaching courses. To date, the assessment of microteaching practices has been dominated by the Classical Test Theory (CTT) approach, which has fundamental limitations in the context of authentic and performance-based assessment. CTT assumes only one source of error (random error) and is unable to simultaneously identify and estimate the contribution of variance from various sources, such as differences between raters, interactions between participants and assessment items, and environmental factors (Monteiro et al., 2019). In many studies, this risks compromising the fairness and accuracy of assessment results, given the variability in rater perceptions and the often high complexity of the teaching situations being assessed.

This study fills this gap by offering a more comprehensive alternative approach, namely through the application of Generalizability Theory (G-Theory). G-Theory allows for a more in-depth reliability analysis by estimating the contribution of each facet to the total score variance. The main novelty of this study lies in the application of a fully crossed p × r × i (person × rater × item) design in the context of microteaching in Indonesia, something that is still very rare as emphasized by Padmadewi & Artini (2019) and Raharjo et al. (2025). This study not only calculates the measurement reliability coefficient empirically through G-Study but also simulates optimal scenarios for assessment improvement through D-Study, so that the results provide a concrete basis for improving the microteaching assessment system. This approach aligns with global trends in educational assessment that emphasize the importance of fairness, precision, and accountability in performance assessment through multi-facet reliability analysis (Andersen et al., 2021; Dzakadzie & Quansah, 2023), while also addressing the need for more scientific and equitable assessment reforms.

Based on this background, the present study aims to analyze the variance structure in microteaching assessment within the Microteaching Course of the Vocational Mechatronics Education Program. Through the G-Theory framework, the study will estimate the contributions of variance from participants, raters, items, and their interactions. Additionally, it will calculate the generalizability coefficient (G-coefficient) and conduct a decision study (D-study) to provide recommendations on the optimal number of raters in the assessment system. The findings are expected to offer practical contributions to the development of microteaching assessment systems in higher education and enrich the methodological discourse on performance-based assessment in Indonesia.

2. METHODS

This study employs a quantitative research approach, with a primary focus on evaluating the reliability of student performance assessments in microteaching using the framework of Generalizability Theory (G-Theory). The study adopts a *fully crossed p × r × i design*, where **p** refers to students (persons), **r** refers to lecturers who evaluate

DOI: https://doi.org/10.17509/e.v24i3.86980
P-ISSN 0852-1190 E-ISSN 2502-0781

the teaching practice (raters), and **i** refers to the observation items. This design enables the identification and estimation of variance components stemming from each facet and their interactions.

Facets refer to the various variables or factors that can influence the measurement results in an assessment, as known in analysis of variance (ANOVA). In the context of educational measurement, facets can include the test taker (person), the rater (rater), the test item (item), the time of day, the place, and even other contextual aspects such as the language of instruction or the instrument format. Each facet has the potential to be a source of variance, whether desirable (because it reflects genuine differences in ability or performance) or undesirable (because it adds to measurement error). Therefore, to obtain accurate estimates of true score variance and error variance, it is important to identify as many relevant facets as possible in the measurement process and classify their contributions appropriately (Society & Group, 2010).

The introduction of the facet concept forms the basis for the application of G-Theory, which allows for reliability analysis by considering the contribution of each facet and their interactions. In this study, this approach was implemented through a fully crossed *p × r × i design*, a measurement structure in which each participant (p = person) is assessed by each rater (r = rater) using each item in the assessment instrument (i = item). In other words, all combinations of participants, raters, and items are fully captured in the collected data. This design allows for a thorough analysis of the variance arising from each facet and the interactions between them, thus providing a more accurate picture of the reliability of the assessment system. In practical terms, this means that: (1) each student is evaluated by both lecturers; and (2) each lecturer provides scores for all items used to observe teaching skills, which are distributed across ten core teaching skill components.

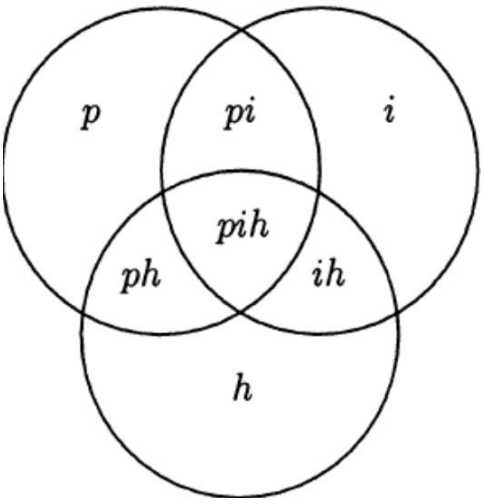

Figure 1. Fully crossed P x R x I design

Figure 1 illustrates the area where all three circles intersect at the center represents the core of the fully crossed design: each student is evaluated by every rater on every item. This highlights that all possible combinations of the three elements, person, rater, and item, are included in the data collection process.

This approach is particularly important in performance assessments such as microteaching, where many non-cognitive factors influence the assessment results, and reliability depends not only on item quality but also on rater consistency and participant characteristics.

DOI: https://doi.org/10.17509/e.v24i3.86980
P-ISSN 0852-1190 E-ISSN 2502-0781

To illustrate the decomposition of observed scores within this fully crossed design, the following equation presents the general linear model structure used in G-Theory:

$$X_{pir} = \mu + (\mu_p - \mu) + (\mu_i - \mu) + (\mu_r - \mu) + (\mu_{pi} - \mu_p - \mu_i + \mu) + (\mu_{pr} - \mu_p - \mu_r + \mu) + (\mu_{ir} - \mu_i - \mu_r + \mu) + (X_{pir} - \mu_{pi} - \mu_{pr} - \mu_{ir} + \mu_p + \mu_i + \mu_r - \mu)$$

Explanation:
μ : grand mean
$\mu_p - \mu$: Person effect (participant)
$\mu_i - \mu$: Item effect
$\mu_r - \mu$: Rater effect
$\mu_{pi} - \mu_p - \mu_i + \mu$: Person × Item interaction
$\mu_{pr} - \mu_p - \mu_r + \mu$: Person × Rater interaction
$\mu_{ir} - \mu_i - \mu_r + \mu$: Item × Rater interaction
$X_{pir} - \mu_{pi} - \mu_{pr} - \mu_{ir} + \mu_p + \mu_i + \mu_r - \mu$: Three way interaction and residual/error

This equation shows how an observed score X_{pir} can be broken down into the grand mean, main effects (student, item, and lecturer), two-way interactions, and a residual term representing unexplained variation or three-way interaction. This modeling framework is fundamental for estimating the contribution of each component to the overall variance in scores.

To identify the sources of variance in the fully crossed p × i × r design, it is necessary to examine the components of variance and their associated expected mean squares (EMS). **Table 1** outlines the sources of variance, the corresponding variance components, and the expected mean squares used in the Generalizability Study (G-study).

Table 1 Variance Components and EMS in the p × i × r Design

Variance Source	Variance Components	Expected Mean Square (EMS)
Person (p)	σ_p^2	$n_i n_r \sigma_p^2 + n_i \sigma_{pi}^2 + n_r \sigma_{pr}^2 + \sigma_{pir}^2$
Item (i)	σ_i^2	$n_p n_r \sigma_i^2 + n_p \sigma_{pi}^2 + n_r \sigma_{ir}^2 + \sigma_{pir}^2$
Rater (r)	σ_r^2	$n_p n_i \sigma_r^2 + n_p \sigma_{pr}^2 + n_i \sigma_{ir}^2 + \sigma_{pir}^2$
Person x Item (pi)	σ_{pi}^2	$n_r \sigma_{pi}^2 + \sigma_{pir}^2$
Person x Rater (pr)	σ_{pr}^2	$n_i \sigma_{pr}^2 + \sigma_{pir}^2$
Item x Rater (ir)	σ_{ir}^2	$n_p \sigma_{ir}^2 + \sigma_{pir}^2$
Person × Item × Rater (pir)	σ_{pir}^2	σ_{pir}^2

DOI: https://doi.org/10.17509/e.v24i3.86980
P-ISSN 0852-1190 E-ISSN 2502-0781

Following the formulation of expected mean squares, **Table 2** presents the formulas used to estimate each variance component based on the mean squares obtained from the analysis of variance (ANOVA). These estimates serve as the basis for determining the contribution of each facet and their interactions to the total measurement variance.

Table 2 Estimation of Variance Components in the p × i × r Design

Variance Source	Estimation of Variance $\sigma^2(\alpha)$
Person (p)	$[MS(p) - MS(pi) - (MS(pr) + MS(pir)]/n_i n_r$
Item (i)	$[MS(i) - MS(pi) - (MS(ir) + MS(pir)]/n_p n_r$
Rater (r)	$[MS(r) - MS(pr) - (MS(ir) + MS(pir)]/n_p n_i$
Person x Item (pi)	$[(MS(pi) + MS(pir)]/n_r$
Person x Rater (pr)	$[(MS(pr) + MS(pir)]/n_i$
Item x Rater (ir)	$[(MS(ir) + MS(pir)]/p$
Person × Item × Rater (pir)	$MS(pir)$

The fully crossed design is essential in G-Theory as it enables a more accurate and comprehensive estimation of all variance components, whether originating from individual students, rater characteristics, item properties, or the interactions among these components. This design allows researchers to obtain in-depth information regarding the sources of score instability in assessments and to conduct a Decision Study (D-study) to estimate score reliability under various configurations of raters and items.

Research Participants

The participants in this study consisted of 32 students enrolled in the Vocational Mechatronics Education Program who were taking the Microteaching course during the 2024/2025 academic year. All participants had completed teaching practice sessions in a microteaching format as part of the program curriculum. Their teaching performance was evaluated by two lecturers who also served as instructors for the Microteaching course.

Research Instrument

This study utilized secondary data derived from observation sheets completed by lecturers responsible for the Microteaching course. In other words, the study did not develop or employ a new assessment instrument directly. Nevertheless, the analyzed data originated from an observation instrument that had been previously designed and implemented by the academic program. The instrument comprised ten observation forms, each corresponding to a specific core teaching skill: (1) lesson introduction, (2) explanation skills, (3) variation in teaching strategies, (4) questioning techniques, (5) facilitating group discussions, (6) teaching small groups and individuals, (7) classroom management, (8) reinforcement techniques, (9) use of media or instructional tools, and (10) lesson closure.

In total, the instrument consisted of 124 rating items using a polytomous Likert-type scale (1–4), developed based on indicators of teaching competence. Each lecturer provided scores for all items for every student.

This study focused its analysis specifically on student teaching skills during microteaching sessions. The assessments were independently conducted by two course lecturers, and the results were tabulated using Excel. A limitation of this study is that the assessments did not include other aspects of the Microteaching course, such as the

DOI: https://doi.org/10.17509/e.v24i3.86980
P-ISSN 0852-1190 E-ISSN 2502-0781

preparation of lesson plans (RPP), social or personality competencies, or out-of-class observations. Therefore, the findings are strictly centered on the reliability and variance structure of teaching performance assessment scores, rather than the full spectrum of course evaluation components.

Data Analysis Technique

Data were analyzed using Generalizability Theory through the EduG 6.1-e software. The analytical procedure consisted of two main stages:

(1) G-Study (Generalizability Study)

This stage was used to estimate the variance components associated with each facet person (p), rater (r), and item (i)as well as their interactions (p × r, p × i, r × i, and the residual p × r × i). These estimates provide insights into the relative contribution of each source to the total score variance. The measurement design specified as P/RI, which denotes a fully crossed design where persons (P) are crossed with a nested combination of raters (R) and items (I). In this context, each person is evaluated across all items and by all raters, which allows for accurate estimation of main effects and interactions across the facets. This design structure is essential for the appropriate decomposition of score variance in microteaching assessments.

(2) D-Study (Decision Study)

In the Decision Study phase, no changes were made to the content or number of items in the assessment instrument. The primary focus was to explore different rater configurations to examine how varying the number of raters would affect the estimation of score reliability and generalizability. Therefore, this study aims to determine the optimal number of raters for the assessment system without modifying the existing instrument.

3. RESULTS AND DISCUSSION

G Study Results

The G coefficient serves as an indicator of the accuracy with which observed scores generalize from a sampled set of behaviors to a broader universe score (Brennan, 2001). Shavelson & Webb (1991) suggested that researchers define their own standards for interpreting G coefficients. One reference commonly cited is from Bracken (1987), who proposed that acceptable reliability levels should be at least 0.80 for subscales and 0.90 for total test scores. Meanwhile, Cicchett (1994) offered interpretive thresholds for reliability coefficients, including Kappa and intraclass correlations, where coefficients below 0.40 are considered poor, 0.40-0.59 as fair, 0.60-0.74 as good, and values above 0.75 as excellent (Parriott, 2016).

Based on the results of this study, the obtained G coefficient was 0.68, which, in general, falls within the moderate category. However, according to Cicchetti's classification, it can be interpreted as indicating good reliability, or at the very least, as acceptable.

Table 3 presents the results of variance component estimation based on Generalizability Theory using a fully crossed p × r × i design. The analysis shows that the person (p) facet contributed the most among the main effects, with a variance estimate of 0.10574, accounting for 24.9% of the total variance. This indicates that individual differences among students substantially influence the observed score variability. In

DOI: https://doi.org/10.17509/e.v24i3.86980
P-ISSN 0852-1190 E-ISSN 2502-0781

contrast, the rater (r) and item (i) facets yielded negative variance estimates (-0.00252 and -0.00084, respectively), each contributing 0.0%. These negative values suggest negligible variation introduced by raters and items, implying a high degree of consistency in scoring and item functioning.

Table 3 Variance Components in Generalizability Theory Analysis

Variance Components	df	SS	MS	Variance	Percentage of Variance (%)
Person (p)	31	1203.54385	38.82400	0.10574	24.9
Rater (r)	1	2.61290	2.61290	-0.00252	0.0
Item (i)	123	24.74042	0.20114	-0.00084	0.0
Person x Rater (pr)	31	390.47581	12.59599	0.09984	23.5
Person x Item (pi)	3813	843.29990	0.22116	0.00238	0.6
Item x Rater (ir)	123	30.79335	0.25035	0.00106	0.2
Person × Item × Rater (pir) (Error)	3813	825.11794	0.21640	0.21640	50.9

The person × rater (pr) interaction accounted for a considerable proportion of the variance, 0.09984 or 23.5%, indicating that student performance varied to some extent depending on which lecturer evaluated them. The person × item (pi) interaction showed a much smaller contribution, with a variance estimate of 0.00238 or 0.6%, suggesting relative stability in student responses across different items. Similarly, the item × rater (ir) interaction contributed minimally to total variance, with an estimate of 0.00106 or 0.2%, indicating consistency in how raters scored across items.

The largest proportion of variance was found in the person × item × rater (pir) interaction, which also represents the residual or error term. This component had a variance estimate of 0.21640, accounting for 50.9% of the total variance. The magnitude of this residual variance underscores the presence of unexplained variability in the assessment system, highlighting potential areas for improvement in reducing measurement error and enhancing the reliability of microteaching evaluations.

D-Study Results

The results of the Decision Study (D-Study) indicate that the reliability of the assessment can be significantly improved by increasing the number of raters in the microteaching evaluation system. In the initial configuration based on the G-Study, which involved two raters, the relative generalizability coefficient (G-coefficient) was 0.68, falling within the moderate category.

Table 4 D-Study Results

No	Option	Number of Raters	Relative G Coefficientf	Relative Standard Error of Measurement
1	G-Study	2	0.67542	0,22541
2	Option 1	3	0.75733	0.18406
3	Option 2	4	0.80621	0.15942
4	Option 3	5	0.83869	0,14260

An increase in the number of raters from two to three successfully improved the G coefficient from 0.68 to 0.76, surpassing the commonly accepted minimum threshold of 0.75 for assessments involving moderate to high-stakes decisions (Brennan, 2001).

DOI: https://doi.org/10.17509/e.v24i3.86980
P-ISSN 0852-1190 E-ISSN 2502-0781

Option 2 and Option 3 demonstrated reliability coefficients that exceeded the optimal threshold of 0.80. In addition to improving reliability, the measurement error was also significantly reduced. The Relative Standard Error of Measurement (Rel. SEM) decreased from 0.22541 in the G-study to only 0.14260 in Option 3. This finding indicates that increasing the number of raters not only enhances reliability but also substantially reduces the uncertainty in students' scores.

The G-Study analysis revealed that the variance component among participants (σ_p^2= 0.1057 or 24.9%) accounted for approximately one-quarter of the total variance, indicating that the instrument is reasonably capable of differentiating students' teaching performance in microteaching. This finding aligns with Brennan (2001) recommendation, which emphasizes that a substantial person variance is a strong indicator of the effectiveness of performance-based measurement. Furthermore, the variance attributable to items and raters was minimal; however, the person × rater interaction (σ_{pr}^2 = 0.0998 or 23.5%) contributed significantly to score bias. Statistically, this suggests that score differences among students were largely influenced by inconsistencies between raters, rather than by actual differences in students' teaching abilities.

Andersen et al. (2021) stated that a high person × rater (p × r) interaction typically arises in performance assessments that rely on subjective observation and lack well-calibrated evaluation criteria among raters. Similarly, a study by Govaerts et al. (2013) found that the p × r variance component can account for more than 30% of total score variance in clinical assessments, reflecting discrepancies in raters' perceptions of participant performance. Empirical evidence shows that rater bias, driven by inter-rater inconsistency, not only increases score error but often emerges when rubrics are not accompanied by adequate rater training and calibration prior to use. As Becker (1999) concluded, "*Generalizability Theory is a promising approach by which rater bias can be studied*," and emphasized that rater calibration is a key strategy for improving score reliability.

Furthermore, Sung et al. (2010) recommended that in performance assessment systems that depend on human judgment, either the number of raters should be increased or rater training should be reinforced to minimize errors stemming from subjectivity and personal interpretation. This recommendation aligns with the findings of Govaerts et al. (2013), who identified the person × rater interaction as a dominant source of measurement error in performance-based assessments. Their study in the field of medical education emphasized the importance of strengthening rater calibration as a means to enhance reliability.

The dominance of the three-way interaction variance component (σ_{pir}^2 = 0.2164 or 50.9%) indicates that nearly half of the total variance originated from the complex interaction between persons, items, and raters. This component represents both systematic and random factors that were not further explored in this study, and it significantly contributed to the reduction in score reliability. Govaerts et al. (2013) found that the P × I × R variance can account for 40-60% of total variance in clinical assessments, and even when instruments and raters are well-prepared, this component can still substantially lower reliability.

In performance-based assessments such as microteaching, an ideal reliability coefficient is typically expected to be ≥ 0.80, particularly when the assessment outcomes are used for summative decisions or other high-stakes evaluations. Such a threshold reflects a high level of score consistency and the assessment system's ability to accurately differentiate students based on their actual abilities. However, the findings of this study revealed a G coefficient of only 0.68, which although classified as moderate

DOI: https://doi.org/10.17509/e.v24i3.86980
P-ISSN 0852-1190 E-ISSN 2502-0781

reliability, falls short of the ideal standard. This value suggests that a considerable proportion of error remains in the assessment system, thereby limiting the trustworthiness of scores when used for formal evaluations or the determination of students' final achievement levels. Therefore, optimization efforts are necessary such as increasing the number of raters or enhancing the quality of inter-rater calibration, to improve the reliability of the assessment system to a level that is academically acceptable.

The D-Study results demonstrate that increasing the number of raters or enhancing inter-rater calibration can substantially reduce the error variance associated with Person × Rater (PR) and Person × Item × Rater (PRI) interactions. This finding is consistent with Hong (2008), Nalbantoğlu Yilmaz & Gelbal (2011), Sung et al. (2010), who reported that adding two to three raters in peer assessment settings increased reliability to ≥ 0.70. Given that the item variance was nearly zero, it can be concluded that the observation instrument is relatively homogeneous and does not contribute to undesirable score variability. This supports the findings of Crawford et al. (2019) in the field of special education, which indicated that rubrics with specific descriptors tend to yield higher reliability by minimizing item-related error (Atilgan, 2019).

The D-Study provides strong justification that the microteaching assessment system would be significantly more reliable if a minimum of three raters were used, and would be optimal with four or five raters. Increasing the number of raters has been shown to reduce the error variance from both the Person × Rater interaction and the residual (PRI) component, which were previously the largest sources of measurement error. This recommendation aligns with the findings of Govaerts et al. (2013), who emphasized that reliability in performance-based assessment depends not only on the instrument, but also on the consistency among raters.

The findings of this study underscore the critical need for policy adjustments at the study program level to enhance the fairness and accuracy of microteaching assessments. Given that reliability substantially increases with the addition of raters and the improvement of inter-rater calibration, the program should consider mandating a minimum of three raters for each microteaching evaluation. Furthermore, structured rater training and calibration workshops should be institutionalized prior to the assessment period to reduce subjective bias and promote scoring consistency. These measures are essential not only to ensure more dependable student performance evaluations but also to uphold the credibility and accountability of the teacher education program in preparing future educators.

4. CONCLUSION

Based on the findings of this study, it can be concluded that the assessment of students' microteaching performance still faces considerable reliability challenges. Although the assessment instrument demonstrated its ability to distinguish between different levels of student performance, the variance analysis revealed that much of the score inconsistency was due to complex interactions among students, raters, and assessment items. Inconsistencies in rater perceptions emerged as a key factor affecting the accuracy of evaluation results, particularly in the absence of a well-calibrated rubric. This indicates that performance-based assessments, such as microteaching, are highly susceptible to measurement error if not designed with a measurement approach that accounts for multiple sources of variance.

Through the Generalizability Theory framework, this study successfully revealed the underlying structure of variance in the assessment system and demonstrated that

DOI: https://doi.org/10.17509/e.v24i3.86980
P-ISSN 0852-1190 E-ISSN 2502-0781

increasing the number of raters is an effective strategy to enhance reliability. Involving more raters in the evaluation process helps reduce inconsistency and improves the accuracy of academic decision-making. To improve the reliability of microteaching performance assessments, the following actions are recommended:

(i) Increase the number of raters to at least three, and ideally four to five, to reduce measurement error and enhance reliability, as demonstrated in the D-Study results.

(ii) Implement regular rater training and calibration sessions to ensure consistency in interpreting assessment criteria and to minimize subjective bias.

(iii) Integrate routine monitoring of reliability using the Generalizability Theory approach to maintain the quality and consistency of the assessment system over time.

(iv) Conduct further research focused on evaluating the quality of the observation rubric, including content validity and the contribution of individual items to score variance, to develop a more comprehensive and robust assessment framework.

5. AUTHORS' NOTE

The authors affirm that there are no competing interests associated with the publication of this manuscript. Furthermore, the authors certify that the content of this paper is entirely original and has not been plagiarized in any form.

6. REFERENCES

Altammar, J., & Aljassar, S. (2021). The effect of microteaching on the enhancement of female mathematics and social studies teachers' skills in the college of education at Kuwait university. *International Journal of Knowledge and Learning*, *14*(4), 324–344. https://doi.org/10.1504/ijkl.2021.118558

Andersen, S. A. W., Nayahangan, L. J., Park, Y. S., & Konge, L. (2021). Use of Generalizability Theory for Exploring Reliability of and Sources of Variance in Assessment of Technical Skills: A Systematic Review and Meta-Analysis. *Academic Medicine*, *96*(11), 1609–1619. https://doi.org/10.1097/ACM.0000000000004150

Aruğaslan, E. (2025). Self-, Peer, and Tutor Assessment in Online Microteaching Practice and Doctoral Students' Opinions. *International Review of Research in Open and Distributed Learning*, *26*(1), 99–117. https://doi.org/10.19173/irrodl.v26i1.7970

Atilgan, H. (2019). Reliability of essay ratings: A study on generalizability theory. *Eurasian Journal of Educational Research*, *2019*(80), 133–150. https://doi.org/10.14689/ejer.2019.80.7

Becker, M. R. (1999). *Rater bias in observational data: A generalizability analysis*.

Bracken, B. A. (1987). Limitations of preschool instruments for identifying children with high intellectual ability: A critical review and recommendations. *Gifted Child Quarterl*.

Brennan, R. L. (2001). Generalizability theory: Statistics for social science and public policy. In *New York: Springer-Verlag.* (Vol. 30).

Brennan, R. L. (2009). Generalizability Theory. In *International Encyclopedia of Education, Third Edition* (pp. 61–68). https://doi.org/10.1016/B978-0-08-

044894-7.00246-3

Brennan, R. L. (2010). Generalizability theory and classical test theory. *Applied Measurement in Education*, *24*(1), 1–21. https://doi.org/10.1080/08957347.2011.532417

Briesch, A. M., Chafouleas, S. M., & Johnson, A. (2016). Use of Generalizability Theory Within K–12 School-Based Assessment: A Critical Review and Analysis of the Empirical Literature. *Applied Measurement in Education*, *29*(2), 83–107. https://doi.org/10.1080/08957347.2016.1138955

Briesch, A. M., Swaminathan, H., Welsh, M., & Chafouleas, S. M. (2014). Generalizability theory: A practical guide to study design, implementation, and interpretation. *Journal of School Psychology*, *52*(1), 13–35. https://doi.org/10.1016/j.jsp.2013.11.008

Cardinet, J., Sandra Johnson, & Pini, G. (2010). *Applying Generalizability Theory using EduG Quantitative Methodology Series*. Routledge.

Cicchett, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*.

Clayson, P. E., & Miller, G. A. (2017). ERP Reliability Analysis (ERA) Toolbox: An open-source toolbox for analyzing the reliability of event-related brain potentials. *International Journal of Psychophysiology*, *111*, 68–79. https://doi.org/10.1016/j.ijpsycho.2016.10.012

Crawford, A. R., Johnson, E. S., Moylan, L. A., & Zheng, Y. (2019). Variance and Reliability in Special Educator Observation Rubrics. *Assessment for Effective Intervention*, *45*(1), 27–37. https://doi.org/10.1177/1534508418781010

Deshpande, S., & Shastri, S. (2020). A cross-sectional study to evaluate teaching skills of postgraduate medical students using component skill approach in microteaching. *Journal of Education and Health Promotion*, *9*(1). https://doi.org/10.4103/jehp.jehp_743_19

Dzakadzie, Y., & Quansah, F. (2023). Modeling unit non-response and validity of online teaching evaluation in higher education using generalizability theory approach. *Frontiers in Psychology*, *14*(September), 1–14. https://doi.org/10.3389/fpsyg.2023.1202896

González-Mélendez, R. C., Sánchez-Rodríguez, M. A., & Robles-López, F. (2023). Validity and reliability of an instrument for assessing microteaching in chemical biological sciences. *Revista Digital de Investigacion En Docencia Universitaria*, *17*(2). https://doi.org/10.19083/ridu.2023.1581

Govaerts, M. J. B., Van de Wiel, M. W. J., Schuwirth, L. W. T., Van der Vleuten, C. P. M., & Muijtjens, A. M. M. (2013). Workplace-based assessment: Raters' performance theories and constructs. *Advances in Health Sciences Education*, *18*(3), 375–396. https://doi.org/10.1007/s10459-012-9376-x

Gudiato, C., Cahyaningtyas, C., & P., N. (2024). Alat Pendeteksi Kebocoran Gas LPG Pada Resto Ayam Bakar dan Goreng Kremes Tata Berbasis Internet Of Things. *G-Tech : Jurnal Teknologi Terapan*, *8*(1), 186–195.

Hendrickson, A., & Yin, P. (2018). Generalizability Theory. In *The Reviewer's Guide to Quantitative Methods in the Social Sciences: Second Edition* (pp. 123–131). https://doi.org/10.4324/9781315755649-9

Hong, S.-J. (2008). Applications of generalizability theory to estimate the variance components in basic nursing technical tests. *Journal of Nursing*, *55*(4), 41–52. https://www.scopus.com/inward/record.uri?eid=2-s2.0-49749131173&partnerID=40&md5=aed8a59f1bf1e930b4a9c9162da6ab45

Jones, E., & Bergin, C. (2019). Evaluating Teacher Effectiveness Using Classroom

Observations: A Rasch Analysis of the Rater Effects of Principals. *Educational Assessment*, *24*(2), 91–118. https://doi.org/10.1080/10627197.2018.1564272

Koech, H. C., & Mwei, P. K. (2019). How secondary school mathematics teachers perceive the effectiveness of microteaching and teaching practice in their preservice education. *Humanities and Social Sciences Letters*, *7*(1), 46–55. https://doi.org/10.18488/journal.73.2019.71.46.55

Komolafe, B. F., Ogunniran, M. O., Zhang, F. Y., & Qian, X. S. (2020). A comparative perspective of teaching skill acquisition in pre-service physics teacher (Pspt) training program in china and nigeria. *Journal of Baltic Science Education*, *19*(3), 356–373. https://doi.org/10.33225/jbse/20.19.356

Matt, G. E., & Sklar, M. (2015). Generalizability Theory. In *International Encyclopedia of the Social & Behavioral Sciences: Second Edition* (pp. 834–838). https://doi.org/10.1016/B978-0-08-097086-8.44027-4

Mishra, R. (2024). Utilizing Online Micro Teaching as the Main Technique in Education Practice. *Proceedings of International Conference on Sustainable Computing and Integrated Communication in Changing Landscape of AI, ICSCAI 2024*. https://doi.org/10.1109/ICSCAI61790.2024.10866844

Monteiro, S., Sullivan, G. M., & Chan, T. M. (2019). Generalizability Theory Made Simple(r): An Introductory Primer to G-Studies. *Journal of Graduate Medical Education*, *11*(4), 365–370. https://doi.org/10.4300/JGME-D-19-00464.1

Nalbantoğlu Yilmaz, F., & Gelbal, S. (2011). Comparison of different designs in accordance with the generalizability theory in communication skills example. *Hacettepe Egitim Dergisi*, *41*, 509–518. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84857386442&partnerID=40&md5=884440963ee930f6c0028796cfd3709f

Özcan, Ö., & Gerçek, C. (2019). Multidimensional analyzing of the microteaching applications in teacher education via videograph. *European Journal of Teacher Education*, *42*(1), 82–97. https://doi.org/10.1080/02619768.2018.1546285

Padmadewi, N. N., & Artini, L. P. (2019). Assessment instruments for improving English teaching skills through microteaching in Indonesia. *Asian EFL Journal*, *21*(2), 49–77. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85063669842&partnerID=40&md5=a0086fa8150e880afe126455909c69e9

Parriott, D. (2016). Using Generalizability Theory to investigate sources of variance of the Autism Diagnostic Observation Schedule-2 with trainees. *UNLV Theses, Dissertations, Professional Papers, and Capstones*, 2721. https://digitalscholarship.unlv.edu/thesesdissertations/2721

Raharjo, H. P., Kusuma, D. W. Y., Mohamed, A. M. D., Rahayu, T., Annas, M., Putra, R. B. A., Suripto, A. W., & Kurniawan, W. R. (2025). Analysis of the online microteaching practice of undergraduate physical education students. *Cakrawala Pendidikan*, *44*(1), 63–71. https://doi.org/10.21831/cp.v44i1.61743

Şen, A. I. (2009). A study on the effectiveness of peer microteaching in a teacher education program. *Egitim ve Bilim*, *34*(151), 165–174. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84886292464&partnerID=40&md5=cd2527f8998e9e397f7aa086fdddffb8

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage Publications.

Society, S., & Group, E. W. (2010). *Edug user guide*.

Sudrajat, A. K., Ibrohim, I., & Susilo, H. (2024). Preservice teachers' reflections on lesson study integration into a microteaching course. *Social Sciences and Humanities Open*, *10*. https://doi.org/10.1016/j.ssaho.2024.101140

DOI: https://doi.org/10.17509/e.v24i3.86980
P-ISSN 0852-1190 E-ISSN 2502-0781

Sung, Y. T., Chang, K. E., Chang, T. H., & Yu, W. C. (2010). How many heads are better than one? The reliability and validity of teenagers' self- and peer assessments. *Journal of Adolescence*, *33*(1), 135–145. https://doi.org/10.1016/j.adolescence.2009.04.004

Teker, G. T., Guler, N., & Uyanik, G. K. (2015). Comparing the effectiveness of spss and edug using different designs for generalizability theory. *Kuram ve Uygulamada Egitim Bilimleri*, *15*(3), 635–645. https://doi.org/10.12738/estp.2015.3.2278

DOI: https://doi.org/10.17509/e.v24i3.86980
P-ISSN 0852-1190 E-ISSN 2502-0781