QUBAHAN ACADEMIC JOURNAL
VOL. 6, NO. 2, May 2026
https://doi.org/10.48161/qaj.v6n2a2462

Optimizing Dependability in Classroom Action Research Skill Assessment: A Multivariate Generalizability Theory Analysis of Alternative Measurement Designs

Roongporn Klyprayong [1], Kamonwan Tangdhanakanond [1] and Sirichai Kanjanawasee [1*]

[1] Department of Educational Research and Psychology, Faculty of Education, Chulalongkorn University, Bangkok 10330, Thailand.

* **Corresponding author:** sirichai.k@chula.ac.th.

ABSTRACT: Reliable assessment of classroom action research skill is essential for supporting valid absolute decisions in teacher education, particularly when performance assessments involve multidimensional constructs and rater-mediated judgment. This study investigated how alternative measurement designs influence the index of dependability and variance structure of classroom action research skill assessment scores within a multivariate generalizability theory (MGT) framework. Specifically, the study compared fully crossed and nested measurement designs and examined the number of raters required to achieve acceptable dependability for absolute decision-making. The participants consisted of 58 fourth-year student teachers majoring in primary education whose classroom action research reports were evaluated by four raters using a multidimensional assessment form aligned with the Plan–Act–Observe–Reflect (PAOR) framework and supported by double-layer scoring rubrics. Data analysis was conducted sequentially using the many-facet Rasch model (MFRM) to examine rater effects, followed by MGT-based generalization and decision studies. The findings showed that the fully crossed design produced a higher composite index of dependability than the nested design (Φ = .8468 vs .7823) and generated different composite universe score variance structures. Under the fully crossed design, three raters were sufficient to achieve acceptable dependability for individual-level absolute decisions, whereas the nested design required four raters to reach a comparable level. The study contributes to educational measurement literature by demonstrating that measurement design influences not only the magnitude of dependability but also the variance structure underlying multidimensional performance assessment scores. The findings further highlight the importance of aligning measurement design with the intended interpretation and use of assessment scores in teacher education contexts.

Keywords: Multivariate generalizability theory, Classroom action research, Performance assessment, Measurement design, Index of dependability.

I. INTRODUCTION

Classroom action research focuses on solving problems that arise in instructional practice and on using research findings to improve student learning and instructional management in a systematic manner. As such, it serves not only as a mechanism for improving classroom practice but also as a means of supporting teachers' ongoing professional development [1-3]. Student teachers, as future teachers, should therefore be developed to possess classroom action research skills so that they are able to analyze instructional problems,

design appropriate interventions, and apply research findings to improve instructional practice with academic justification [4-5].

Theoretically, classroom action research is grounded in Lewin's [6] concept of action research and was later developed by Kemmis and McTaggart [7] into a cyclical process that can be implemented repeatedly and continuously. This cycle consists of four iterative stages: Plan, Act, Observe, and Reflect (PAOR). These characteristics indicate that classroom action research skill is process-oriented and inherently multidimensional because each stage reflects distinct but related behavioral indicators. In practice, these dimensions are expressed through the writing of a classroom action research report, beginning with problem identification and research design, followed by research implementation in accordance with ethical principles and research methodology, data collection and analysis, and concluding with reflection on the application of research findings [8-9]. Accordingly, the assessment of classroom action research skill should reflect its overall multidimensional structure rather than treating each dimension as entirely separate from the others.

Figure 1 illustrates the conceptual framework of the present study and explicitly represents the PAOR dimensional structure of classroom action research skill, comprising Plan, Act, Observe, and Reflect as related but distinct dimensions within the construct. It also shows how these dimensions were operationalized through classroom action research reports, items, behavioral indicators, and a double-layer scoring rubric, as well as how the resulting scores were evaluated through MFRM and MGT.

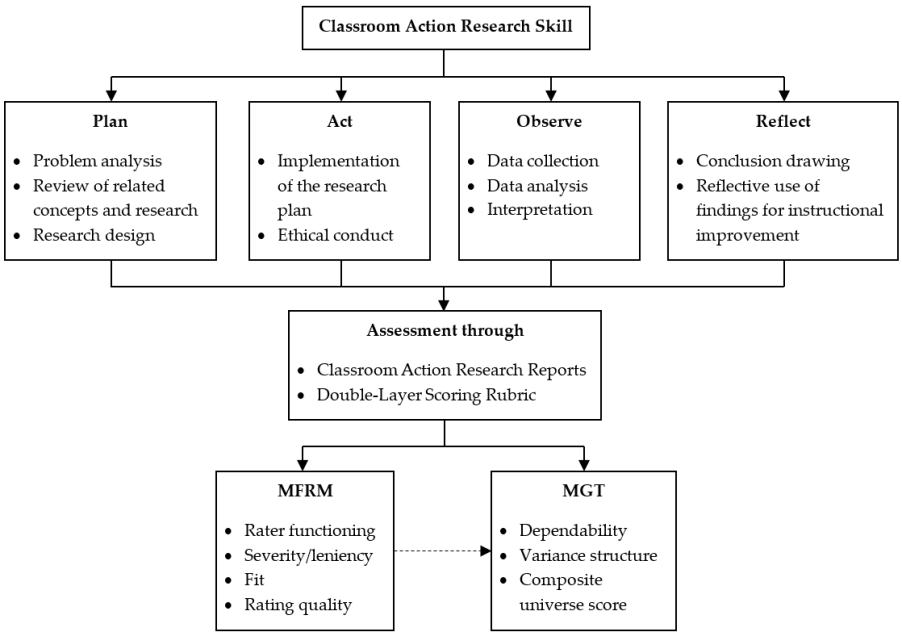


FIGURE 1. Conceptual framework and PAOR dimensional structure of classroom action research skill assessment.

Given these characteristics, classroom action research skill assessment should be based on authentic performance and criterion-referenced scoring procedures, particularly scoring rubrics that support clear and meaningful score interpretation. Research on performance assessment has consistently shown that scoring rubrics help make the quality of complex skills more visible and interpretable [10-12]. Recent performance assessment studies have also shown that the choice of scoring format may influence score reliability, rater consistency, and the usefulness of assessment results for practical decision-making [13]. However, even when scoring rubrics are used, rater-related measurement error remains a major concern, especially in

assessments involving multiple raters and subjective judgment [14]. Recent studies continue to show that variation among raters can affect score quality, particularly when raters differ in experience, interpretation of scoring criteria, or severity and leniency in judgment [15]. Recent work also indicates that rater effects may introduce construct-irrelevant variance and influence the fairness and validity of performance assessment scores [16-17]. These findings suggest that observed scores may reflect not only learners' actual performance but also characteristics associated with the raters and the assessment context.

Beyond consistency and the index of dependability, the interpretation and use of performance assessment scores also require a broader validity perspective [18-19]. Messick's unified view of validity emphasizes that score meaning should be supported by multiple sources of evidence, including the substantive basis of the construct, the structural properties of the assessment, and the implications of score use. Similarly, Kane's argument-based approach highlights the need to justify the inferences drawn from observed scores to intended decisions. In the present context, these perspectives are particularly relevant because classroom action research skill assessment involves multidimensional performance, rater-mediated judgment, and criterion-referenced interpretation. Accordingly, evidence regarding rater functioning, variance structure, and the index of dependability is important not only for score consistency but also for supporting the interpretation and use of scores for absolute decisions.

In this regard, recent work in educational measurement has increasingly emphasized that score quality in performance assessment depends not only on the quality of the rubric or task but also on the underlying measurement design, including the number of raters, the number of items, and the arrangement of scoring facets. Recent studies further suggest that the number and configuration of raters can materially affect the reliability of performance scores and the defensibility of score-based decisions [13, 20]. This issue is especially important when scores are used for absolute decisions, because such decisions require a sufficient index of dependability for individual-level interpretation. Nunnally and Bernstein [21] noted that assessments involving human judgment must carefully consider multiple sources of variation; otherwise, obtained scores may not adequately represent learners' actual ability.

Within the framework of Generalizability Theory (GT), the rater can be treated as a facet that contributes to score variation alongside person and item facets [22-24]. GT enables the systematic decomposition of variance into multiple sources and their interactions, thereby providing a stronger basis for evaluating the index of dependability under different measurement designs. Brennan [22-23] explained that design structures such as fully crossed and nested designs directly affect variance-component estimates, universe score variance, and the index of dependability (**Φ**), particularly for absolute decisions. Recent GT-based studies in performance assessment have continued to demonstrate the usefulness of GT for examining rater, task, and design-related sources of score variability in contexts where scores support high-stakes or classification-based decisions [20]. Thus, decisions regarding the number of raters and the arrangement of raters across scoring conditions are psychometric issues that are directly related to the quality and intended use of assessment scores.

When the skill being assessed has a multidimensional structure, multivariate generalizability theory (MGT) provides an especially useful framework because it allows the relationships among dimensions to be examined simultaneously and makes it possible to evaluate the variance structure of composite universe scores. Brennan [22-23] pointed out that even when dimensions are assigned equal nominal weights, the composite universe score variance may still differ, which in turn affects the interpretation of the index of dependability and the appropriateness of the measurement design. Recent applications of MGT have demonstrated its value for evaluating multidimensional scores, composite scores, and alternative scoring methods in complex assessment contexts [25-26]. In addition, recent methodological work indicates that MGT continues to develop in response to increasingly complex assessment designs [27]. This is particularly relevant in the assessment of classroom action research skill, where the PAOR dimensions are conceptually related but not interchangeable.

However, prior GT-based studies have several limitations in relation to the present study. Many have focused on overall score dependability or separate dimension-level estimates rather than on how alternative measurement designs affect the variance structure of multidimensional scores [20, 25]. Even in more recent MGT applications, less attention has been given to comparing design structures such as fully crossed and

nested arrangements in rater-mediated contexts where scores are used for absolute decisions [25-27]. These limitations are especially important in the present context because both dependability and the structure of the composite universe score may be sensitive to the measurement design.

In the present study, the use of MGT was necessary rather than optional. A univariate GT approach with separate sub scores would have allowed the dependability of each PAOR dimension to be examined independently, but it would not have captured the covariance structure among dimensions or the variance structure of the composite universe score. Because classroom action research skill was conceptualized in this study as a multidimensional construct composed of related but non-interchangeable dimensions, and because the study aimed to compare how alternative measurement designs affect both dependability and the structure of composite scores, MGT provided a more appropriate framework than separate univariate analyses. In this sense, MGT made it possible to evaluate not only the quality of dimension-specific scores, but also how the PAOR dimensions functioned together under each measurement design.

Alternative psychometric frameworks may also be used to examine score quality in complex performance assessment. For example, IRT-based approaches can provide useful information about rating processes and rater-related functioning in rater-mediated assessments, whereas Bayesian generalizability approaches offer additional flexibility for estimating variance components under complex designs. However, these alternatives were not the primary focus of the present study because the research questions centered on multidimensional score interpretation, composite universe score variance, and the comparison of alternative measurement designs for absolute decisions. In this context, the combination of MFRM and MGT was considered the most directly relevant framework for the aims of the study [28-29]. At the same time, in rater-mediated assessment, it is important to examine rater effects before conducting MGT analysis. For this reason, the present study applied the many-facet Rasch model (MFRM) to examine rater effects in terms of scoring severity, leniency, and diagnostic information about rating quality [30-31], prior to the MGT analysis. This sequencing is consistent with recent work showing that MFRM can provide important diagnostic evidence about rater functioning and scoring quality before subsequent score interpretation and design evaluation [17, 32].

Although recent studies in performance assessment and generalizability-based research have highlighted the importance of multidimensional score interpretation, rater effects, and design-related sources of measurement error, there remains limited evidence on how alternative measurement designs influence the index of dependability and variance structure of multidimensional performance assessment scores, particularly when those scores are used for absolute decisions [13, 20, 25]. This limitation is especially evident in studies involving process-oriented skills assessed through multiple raters and composite scores across related dimensions. Moreover, existing applications of MGT have primarily examined general score dependability or particular scoring methods rather than comparing how fully crossed and nested designs may produce different variance structures and different rater requirements [25-26]. Therefore, this study sought to examine the index of dependability (Φ) of student teachers' classroom action research skill assessment scores under different measurement designs within the multivariate generalizability theory framework and to compare dependability across different numbers of raters in order to identify a design that is appropriate for absolute decisions.

II. THEORETICAL FRAMEWORK

Classroom action research skill can be conceptualized as a multidimensional construct grounded in the PAOR cycle of action research: Plan, Act, Observe, and Reflect. These four dimensions represent interrelated but distinct components of classroom inquiry. In the present study, the Plan dimension involves problem analysis, review of relevant concepts and research, and classroom research design; the Act dimension concerns implementation of the research plan and ethical conduct; the Observe dimension concerns data collection, analysis, and interpretation; and the Reflect dimension concerns conclusion drawing and reflective use of findings for instructional improvement. This conceptualization is consistent with the view that classroom action research skill is process-oriented and should be assessed as a multidimensional performance rather than as a single undifferentiated trait. Because the dimensions are conceptually related

but not interchangeable, score interpretation should take into account both dimension-specific performance and the meaning of their combination within the broader classroom action research process [6-7, 22]. This multidimensional view also aligns with the PAOR-based structure used in the present assessment framework.

In the present study, the PAOR model serves as the substantive framework for defining classroom action research skill, whereas measurement theory provides the basis for evaluating how well that construct is represented in observed scores. More specifically, the PAOR dimensions were operationalized through assessment items, behavioral indicators, and a double-layer scoring rubric, thereby translating the construct into an assessable multidimensional performance structure. From this perspective, measurement theory is not separate from the PAOR model, but integrative: MFRM was used to examine the functioning of rater-mediated scoring, and MGT was used to evaluate the dependability and variance structure of the resulting multidimensional scores. Thus, the present study links the PAOR model to measurement theory by treating the PAOR dimensions as the substantive basis of score interpretation and MFRM/MGT as the psychometric basis for evaluating the quality of that interpretation [7, 18-19, 22].

At the same time, classroom action research skill was assessed through classroom action research reports using a double-layer scoring rubric, making it a rater-mediated performance assessment rather than a purely objective test. The assessment was designed to evaluate both process and product aspects of performance across the PAOR dimensions, and the rubric format was selected because it provides detailed and systematic evidence at the dimension level. Under such conditions, observed scores may be influenced not only by student teachers' actual performance but also by raters' interpretations of rubric criteria. For this reason, rater effects are theoretically relevant rather than incidental to the assessment process. In performance assessment, variation in scoring severity, leniency, or use of criteria may introduce construct-irrelevant variation into observed scores, which in turn affects the appropriateness of score interpretation and use. Accordingly, the present study treats classroom action research skill assessment as a multidimensional, rubric-based, and rater-mediated assessment context in which score quality depends on both construct representation and scoring quality [10, 33-34].

Within this context, Generalizability Theory provides the overarching psychometric framework for evaluating score quality under multiple sources of measurement error. Unlike classical test theory, which treats error as a single undifferentiated component, Generalizability Theory decomposes score variation into multiple facets and their interactions, thereby making it possible to examine how particular measurement conditions influence score dependability. This feature is especially important in the present study because classroom action research skill was evaluated under conditions involving both items and raters. Moreover, because the construct consists of four related dimensions, Multivariate Generalizability Theory is more appropriate than univariate GT. Multivariate GT allows the relationships among dimensions to be examined simultaneously and provides a basis for evaluating the variance structure of composite universe scores rather than relying only on separate dimension-level estimates. As Brennan [22] noted, even when dimensions are assigned equal nominal weights, the composite universe score variance may still differ, which has implications for the interpretation of the index of dependability and the appropriateness of alternative measurement designs. Recent applications have also demonstrated the value of MGT for evaluating multidimensional scores, composite scores, and alternative scoring methods in complex assessment contexts [25, 35]. In the present study, this framework is directly relevant because the design comparison focuses on how different arrangements of raters and items affect the index of dependability for absolute decisions, as well as the variance structure underlying composite scores [22, 24].

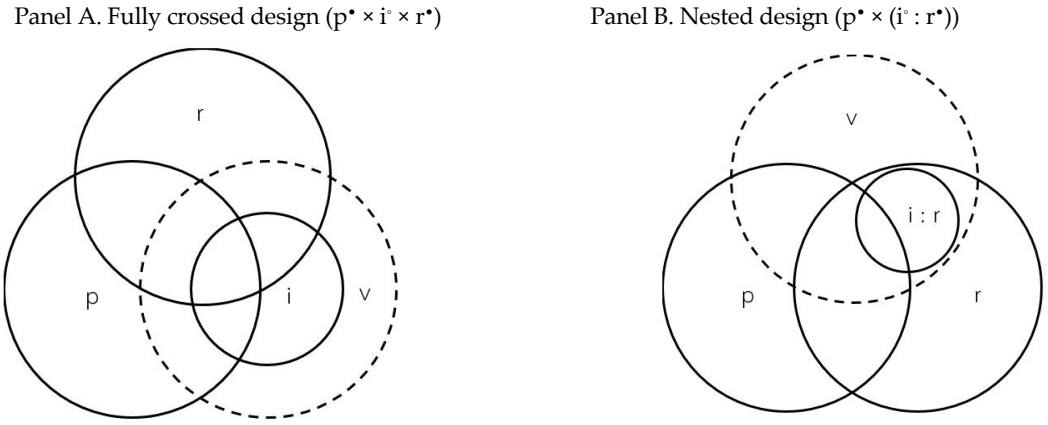
Although Multivariate Generalizability Theory serves as the principal framework for evaluating variance structure, composite universe scores, and the index of dependability, examination of rater functioning remains necessary before those estimates can be interpreted. For this reason, the many-facet Rasch model is positioned in the present study as a complementary analytic approach for examining rater effects, particularly severity, leniency, and diagnostic information about rating quality. This role is consistent with recent work showing that MFRM and Generalizability Theory can be used in complementary ways to investigate rater effects and support score interpretation in rater-mediated assessments [30]. The role of MFRM in the present framework is not to replace MGT, but to clarify whether the observed ratings

functioned as intended before MGT is used to evaluate design-dependent score quality. In the present study, the results of the MFRM analysis were not integrated quantitatively into the MGT analysis because the two approaches served different analytic purposes. MFRM was used diagnostically to examine rater functioning, whereas MGT was used to evaluate design-dependent dependability and variance structure. Accordingly, MFRM informed the interpretation of the observed ratings prior to MGT, rather than serving as a source of adjusted scores for the multivariate generalizability analysis. This sequencing is theoretically justified because, in rater-mediated performance assessment, evidence about rater functioning supports the interpretation of observed scores prior to the estimation of their dependability under alternative measurement designs. Thus, the present study integrates the PAOR-based multidimensional structure of classroom action research skill, the rater-mediated nature of the scoring process, and the complementary roles of MFRM and MGT in supporting score interpretation and use for absolute decisions. From a broader validity perspective, this integration is also consistent with the view that score interpretation should be supported by evidence concerning construct representation, scoring processes, and the intended use of scores [18-19].

III. MATERIAL AND METHOD

1. RESEARCH DESIGN

This study employed a quantitative research design to examine the index of dependability of classroom action research skill assessment scores under alternative measurement designs within the multivariate generalizability theory (MGT) framework. The assessment was conceptualized as a multidimensional performance assessment reflecting four dimensions of the classroom action research process: Plan, Act, Observe, and Reflect (PAOR). Because the intended interpretation of scores was criterion-referenced, the study focused on the index of dependability (**Φ**) for absolute decisions. The analysis proceeded in two stages. First, the many-facet Rasch model (MFRM) was used to examine rater effects in the scoring process. Second, the score data were analyzed within the MGT framework through a generalization study (G-study) and a decision study (D-study) to estimate variance and covariance components, the variance structure of multidimensional scores, and the index of dependability under alternative measurement designs.



p• = student teachers; i◦ = fixed items; r• = raters as a random facet; v = fixed PAOR dimensions

FIGURE 2. Conceptual diagrams of the multivariate generalizability designs examined in the study.

Two measurement designs were examined. The first was a fully crossed design (p• × i◦ × r•), in which all student teachers were assessed on all fixed items by all raters. The second was a nested design (p• × (i◦ : r•)), in which different sets of raters were assigned to different dimensions of the assessment. These multivariate

generalizability designs are illustrated in Figure 2. Panel A presents the fully crossed design ($p\bullet \times i\circ \times r\bullet$), in which all student teachers were assessed on all fixed items by all raters. Panel B presents the nested design ($p\bullet \times (i\circ : r\bullet)$), in which different rater sets were assigned to different PAOR dimensions. In both diagrams, $p\bullet$ denotes student teachers, $i\circ$ denotes fixed items, $r\bullet$ denotes raters treated as a random facet, and v denotes the fixed PAOR dimensions underlying the multidimensional assessment structure.

2. PARTICIPANTS

The participants were 58 fourth-year student teachers enrolled in the Bachelor of Education program in Primary Education at the Faculty of Education, Ramkhamhaeng University, during the second semester of the 2024 academic year. The sample was selected by stratified random sampling based on the subject area of the teaching practicum and classified into two groups, Science–Mathematics and Language–Social Sciences, in order to ensure diversity in instructional context. The required sample size was estimated using G*Power at a significance level of .05 and statistical power of .95, yielding a minimum required sample of 52 participants. To reduce the risk of data attrition, the sample size was increased by approximately 10%, resulting in a final sample of 58 student teachers.

In this study, the primary object of measurement was the student teacher (person facet; p). The classroom action research reports were scored by four raters, all of whom had prior experience in evaluating classroom action research and at least five years of teaching experience in higher education. The use of four raters was based on both practical and methodological considerations. Practically, four raters represented the maximum feasible scoring condition within the study context while allowing all reports to be scored under controlled conditions. Methodologically, the study did not assume in advance that four raters would be universally sufficient for all MGT applications. Rather, four raters were used as the maximum observed rater condition in the G-study so that the D-study could estimate the index of dependability under reduced and alternative numbers of raters. In this sense, the adequacy of the number of raters was treated as an empirical question to be examined through the D-study rather than as a fixed assumption.

3. INSTRUMENT

The instrument used in this study was a classroom action research skill assessment form developed to reflect the multidimensional structure of classroom action research skill based on the PAOR process and supported by double-layer scoring rubrics. The instrument was developed previously, and its quality had been examined in earlier studies [36-37].

The instrument was developed systematically from the PAOR framework of classroom action research, covering four dimensions: Plan, Act, Observe, and Reflect. The first layer of the double-layer scoring rubric was constructed by defining three performance levels for each assessment item (3 = Good, 2 = Satisfactory, 1 = Needs improvement), together with performance descriptors aligned with the behavioral indicators of each dimension. Content validity, language appropriateness, and the suitability of the scoring criteria were then examined by seven experts, including four content specialists and three specialists in educational measurement and evaluation. The quality criteria followed established CVI guidelines, with I-CVI expected to be at least .80 and S-CVI above .90. The second layer of the rubric was developed by establishing cut scores through the extended Angoff method, using eight experts with at least five years of experience in evaluating classroom action research reports. Prior to cut-score judgment, the experts participated in a workshop and reviewed the assessment context and sample reports to support informed judgments.

The assessment form consisted of 10 items with a total of 23 behavioral indicators covering the four dimensions of classroom action research: Plan, Act, Observe, and Reflect. The first layer of the rubric provided behavioral descriptions for each item and defined three score levels: 3 = Good, 2 = Satisfactory, and 1 = Needs improvement. The second layer involved the determination of cut scores to interpret the total scores from Layer 1 into three levels of skill quality using the extended Angoff method: 3 = Good, 2 = Satisfactory, and 1 = Needs improvement. The use of the double-layer scoring rubric supported detailed and systematic scoring across dimensions, reduced ambiguity in judgment, and increased scoring consistency [34, 38-39]. In addition, the cut scores in Layer 2 enhanced the clarity and appropriateness of total score interpretation for absolute decisions. An example of the rubric is provided in the Appendix.

For the purposes of MGT, the items were treated as a fixed facet rather than as a random facet. This decision was based on the fact that the items were not intended to represent a random sample from a broad and interchangeable universe of classroom action research items. Instead, they were purposefully developed to operationalize the predefined PAOR dimensions and their associated behavioral indicators. Accordingly, score interpretation in this study was intended to generalize to the specified assessment structure and rubric used in the study, rather than to an unlimited universe of alternative items.

The instrument had prior empirical evidence supporting both validity and reliability. Evidence of content validity was supported by I-CVI values ranging from 0.86 to 1.00 and an S-CVI of 0.98. Additional support for criterion-related validity was obtained from evidence based on relationships with other variables at both scoring layers. At Layer 1, the total score showed a strong correlation with the classroom action research course examination score (r = .942), whereas at Layer 2, the quality level showed a moderate correlation with the classroom action research course grade (rs = .531). Regarding reliability, high levels of intra-rater reliability and inter-rater reliability were found (.991 and .889, respectively), and substantial agreement among raters was indicated by an intraclass correlation coefficient of .930. These properties support the use of the instrument for examining the index of dependability (**Φ**) in the present study.

4. PROCEDURE

Data were collected from the classroom action research reports produced by the student teachers and used as empirical evidence for classroom action research skill assessment. Before scoring, all reports were anonymized to remove identifying information and reduce potential bias. The reports were then randomly ordered, and the same scoring sequence was distributed to all raters in order to control the scoring condition. Before operational scoring began, the raters participated in a calibration session to establish a shared understanding of the scoring criteria and the use of the double-layer scoring rubric. Practice scoring was conducted using sample reports so that the raters could discuss the interpretation of descriptors and reduce avoidable scoring discrepancies. After calibration, the raters scored the reports independently over a period of two months, and all scores were recorded through an online form system for subsequent analysis.

Ethical approval for this study was granted by the Research Ethics Review Committee for Research Involving Human Subjects (COA No. 198/68; Project No. 680154). The project underwent expedited review, and the approval remained valid from May 13, 2025, to May 12, 2026. Informed consent was obtained from all participants prior to data collection.

5. DATA ANALYSIS

Data analysis was conducted sequentially, beginning with the examination of rater effects through the many-facet Rasch model (MFRM). The multivariate index of dependability (**Φ**) was then analyzed through both a generalization study (G-study) and a decision study (D-study).

5.1 Many-facet Rasch Model (MFRM)

Rater effects were examined using R with the TAM package under a 1PL many-facet Rasch model. The modeled facets were student teachers, items, and raters, allowing the estimation of rater severity and leniency in logit units. These estimates were used to examine differences among raters before the score data were analyzed for the index of dependability within the MGT framework.

5.2 Multivariate Index of Dependability Analysis

Within the MGT framework, the universe of admissible observations was specified as follows. Student teachers were treated as a random facet (p•), because they represented a sample from the broader population of student teachers to whom interpretation was intended to generalize. Raters were also treated as a random facet (r•), assuming that they represented a sample from a population of raters with comparable qualifications and expertise in assessing classroom action research skills. The items within each PAOR dimension were treated as a fixed facet (i◦), because they were intentionally developed from the conceptual framework of classroom action research skill and were not sampled from a broad interchangeable item universe.

5.3 Generalization Study (G-Study)

In the G-study, variance and covariance components were estimated under two designs: (1) a fully crossed design ($p\bullet \times i\circ \times r\bullet$), in which all 58 student teachers were assessed on all items across the four PAOR dimensions by all four raters; and (2) a nested design ($p\bullet \times (i\circ : r\bullet)$), in which all 58 student teachers were assessed by different sets of raters, with two raters per set. For the nested design, raters were first assigned to rater sets by simple random assignment prior to assessment. The two rater sets were then assigned to grouped PAOR dimensions, with the first set assessing the Plan and Act dimensions and the second set assessing the Observe and Reflect dimensions. This grouping was intended to reflect the process structure of classroom action research, with Plan and Act representing earlier process phases and Observe and Reflect representing later evidence-based and reflective phases. Accordingly, the nested arrangement was intended to reflect both procedural control and substantive coherence. However, because only one nested configuration was examined in the present study, the findings from this design should be interpreted within that specific arrangement. In the multivariate analysis, variance and covariance components among the PAOR dimensions were estimated to derive the composite universe score under the multidimensional structure.

In the D-study, the number of raters was varied from 1 to 4 in order to compare the index of dependability ($\mathbf{\Phi}$) under each measurement design using mGENOVA version 2.1. Because the assessment results were intended for absolute decisions, the criterion for acceptable dependability was set at $\mathbf{\Phi} \geq .80$. Although Generalizability Theory does not prescribe a universal cutoff criterion, coefficients of .80 or higher are commonly regarded as adequate for individual-level decision-making in educational measurement [21]. The D-study therefore allowed the adequacy of alternative rater conditions to be evaluated empirically under both the fully crossed and nested designs.

5.4 Assumptions of the Measurement Models

a. Assumptions of MFRM

The MFRM analysis assumed that the observed ratings could be modeled as a function of student teacher performance, item difficulty, and rater severity within a common scoring structure. It was further assumed that raters scored independently, that the rating scale categories functioned in an ordered manner, and that systematic rater differences could be modeled explicitly rather than treated as undifferentiated measurement error [40-41].

b. Assumptions of GT/MGT

The GT/MGT analysis assumed that classroom action research skill was sufficiently stable during the scoring period, that observed score variation could be decomposed into interpretable facet-related sources of variance, and that the specified design represented the intended universe of admissible observations. Because the study focused on absolute decisions, the D-study emphasized absolute error and the estimation of the index of dependability ($\mathbf{\Phi}$) rather than coefficients intended for relative decisions [35, 42-44].

IV. RESULTS

1. RESULTS OF THE MANY-FACET RASCH MODEL (MFRM)

The MFRM results presented in Table 1 suggest that variation in rater severity and leniency was low. The severity/leniency estimates ranged narrowly from –.0554 to .0618 logits, reflecting the logic of the many-facet Rasch model, in which relative differences among raters are examined in logit units to evaluate rater effects on scores [40-41]. Rater 1 showed the highest level of severity (.0618 logits), whereas Rater 3 showed the greatest leniency (–.0554 logits). Overall, the narrow spread of these estimates suggests that the four raters scored in a broadly comparable manner.

In addition, rater fit statistics were examined through outfit and infit mean-square indices. These values indicated an overfit pattern, suggesting that the raters' scoring was more predictable than expected by the model rather than showing problematic misfit. This pattern should be interpreted cautiously, as it may also

reflect highly homogeneous scoring behavior or limited variability in the observed ratings. The model-based EAP reliability coefficient was .91, indicating a high level of score consistency within the fitted MFRM framework. The estimated step parameters were also ordered monotonically, providing preliminary evidence that the rating scale categories functioned in the intended ordinal direction. Taken together, these results suggest that the ratings showed a generally stable scoring pattern across raters and provide preliminary diagnostic support for using the ratings in the subsequent multivariate generalizability analysis [30-31, 40-41].

Table 1. Rater severity estimates and fit statistics from the many-facet Rasch model.

Raters	Severity/Leniency (Logit)	SE (Standard error)	Outfit MSQ	Infit MSQ
1	.0618	.048	.0268	.0243
2	-.0032	.048	.0273	.0282
3	-.0554	.049	.0322	.0296
4	-.0030	.084	.0261	.0254

Note. The model-based EAP reliability coefficient was .91. The estimated step parameters were ordered monotonically (step1 = –3.577, step2 = 0.436, step3 = 3.141), providing preliminary evidence that the rating scale categories functioned in the intended ordinal direction.

2. RESULTS OF THE MULTIVARIATE INDEX OF DEPENDABILITY ANALYSIS

2.1 Results of the Generalization Study (G-Study)

Table 2 shows that under the fully crossed design, the largest proportion of variance in most dimensions was attributable to the person facet, particularly in Plan, Observe, and Reflect. This pattern indicates that score variation under this design was driven primarily by differences among student teachers rather than by unwanted facet-related sources of error. By contrast, variance associated with raters and interaction components was generally smaller, although some interaction-related variation remained in specific dimensions. Overall, this variance structure is consistent with the relatively higher dependability coefficients obtained under the fully crossed design. Because the items were treated as a fixed facet in the present study, the item-related components were not interpreted in the same way as random-facet variance components. Instead, they were understood as part of the specified assessment structure defined by the PAOR-based rubric.

Table 2. Variance (in bold) and covariance components for the fully crossed design ($p^{\bullet} \times i^{\circ} \times r^{\bullet}$).

Source of variation	Estimated variance components							
	Plan	**%**	**Act**	**%**	**Observe**	**%**	**Reflect**	**%**
Person (p)	.0780	80.83	.3685		.1172		.0000	
	.0135		.0171	50.89	.0462		.2773	
	.0060		.0011		.0340	77.98	.5831	
	.0000		.0100		.0295		.0755	79.39
Item (i)	.0000	0.00						
			.0000	0.00				
					.0000	0.00		
							.0000	0.00
Rater (r)	.0000	0.00						
	.0000		.0078	23.21				
	.0000		.0000		.0000	0.00		
	.0000		.0000		.0000		.0012	1.26
pi	.0000	0.00						
			.0000	0.00				
					.0000	0.00		
							.0000	0.00

Source of variation	Estimated variance components							
	Plan	%	Act	%	Observe	%	Reflect	%
pr	.0000	0.00						
	.0002		.0086	25.60				
	.0000		.0000		.0000	0.00		
	.0001		.0003		.0000		.0184	19.35
ir	.0039	4.04						
			.0000	0.00				
					.0012	2.75		
							.0000	0.00
pir,e	.0146	15.13						
			.0001	0.30				
					.0084	19.27		
							.0000	0.00
Total	.0965	100.00	.0336	100.00	.0436	100.00	.0951	100.00

Table 3 shows that the variance structure changed under the nested design. Compared with the fully crossed design, a smaller proportion of variance was attributable to the person facet in some dimensions, whereas larger proportions were associated with rater-related and residual components. This shift was especially evident in the Act dimension, where person variance was negligible and larger proportions of variance were associated with rater and person-by-rater-related sources. This pattern helps explain why the nested design produced lower dependability coefficients overall and an extremely low dependability estimate for the Act dimension in particular.

Table 3. Variance (in bold) and covariance components for the nested design ($p^{\bullet} \times (i^{\circ} : r^{\bullet})$).

Source of variation	Estimated variance components							
	Plan	%	Act	%	Observe	%	Reflect	%
Person (p)	.0383	18.15	21.6083		.2489		.2601	
	.0260		.0000	0.00	2.3297		7.9394	
	.0069		.0020		.0201	19.98	.9405	
	.0101		.0097		.0264		.0391	25.08
Rater (r)	.0167	7.91						
	.0000		.0459	46.84				
	.0052		.0000		.0000	0.00		
	.0159		.0000		.0031		.0076	4.87
i:r	.0219	10.38						
			.0000	0.00				
					.0112	11.13		
							.0000	0.00
pr	.0292	13.84						
	.0000		.0510	52.04				
	.0000		.0000		.0000	0.00		
	.0000		.0017		.0065		.1092	70.05
pi:r,e	.1049	49.72						
			.0011	1.12				
					.0693	68.89		
							.0000	0.00
Total	.2110	100.00	.0980	100.00	.1006	100.00	.1559	100.00

The interaction-related components also provide additional insight into the score structure. In the present context, the person-by-rater interaction ($p\bullet \times r\bullet$) reflects the extent to which student teachers were not scored in exactly the same relative way by different raters, beyond average differences in rater severity. The person-by-item interaction ($p\bullet \times i\circ$) in the fully crossed design reflects the extent to which student teachers' performance varied across assessment items, suggesting that strengths and weaknesses were not distributed uniformly across the behavioral indicators. More broadly, these interaction-related components are important because they indicate that score variation was shaped not only by the main effects of student teachers, raters, or items, but also by inconsistency in how performance was expressed across items and interpreted across raters. In the present study, such components were more pronounced in some dimensions and design conditions than in others, which helps explain differences in the resulting dependability estimates.

The variance components can also be interpreted in terms of error variance decomposition. In the present study, the index of dependability depended on the extent to which total score variance was attributable to the universe score rather than to unwanted sources of measurement error. Under the fully crossed design, the error structure was comparatively more favorable because a larger proportion of variance in most dimensions was attributable to student teachers, whereas rater-related and interaction-related components were generally smaller. Under the nested design, however, a greater proportion of variance in some dimensions was associated with rater-related and residual components, indicating a less favorable error structure for dependable score interpretation. In this sense, the lower dependability coefficients under the nested design can be understood not merely as lower coefficients, but as the result of a larger proportion of error-related variance relative to universe score variance. This pattern was especially evident in the Act dimension under the nested design, where the balance between universe score variance and error-related variance was particularly unfavorable.

As presented in Table 4, the fully crossed design ($p\bullet \times i\circ \times r\bullet$) yielded a composite index of dependability ($\mathbf{\Phi}$) of .8468, which exceeded the .80 criterion used in the present study for individual-level absolute decisions. In contrast, the nested design ($p\bullet \times (i\circ : r\bullet)$) produced a composite $\mathbf{\Phi}$ of .7823, which was close to this criterion but remained slightly below it. At the dimension level, the fully crossed design yielded higher $\mathbf{\Phi}$ values than the nested design across all PAOR dimensions. This difference was especially notable for the Act dimension, for which the nested design produced an extremely low dependability coefficient ($\mathbf{\Phi}$ = .0008).

Table 4 also shows that the relative contribution of each dimension to the composite universe score differed by measurement design. Under the fully crossed design, the Plan dimension contributed the largest proportion (39.89%), followed by Observe (26.10%), Reflect (24.61%), and Act (9.40%). Under the nested design, the corresponding proportions were 37.61%, 12.49%, 24.93%, and 24.97%, respectively. This pattern indicates that the structure of the composite universe score was not invariant across the two designs. It should also be noted that a low dependability coefficient at the dimension level does not necessarily imply a negligible statistical contribution to the composite universe score. Rather, these two results reflect different aspects of multidimensional score interpretation.

Table 4. Index of dependability for each PAOR dimension under the $p\bullet \times i\circ \times r\bullet$ and $p\bullet \times (i\circ : r\bullet)$ designs.

PAOR	$p\bullet \times i\circ \times r\bullet$		$p\bullet \times (i\circ : r\bullet)$	
	$\mathbf{\Phi}$	Comp Univ Score Var (%)	$\mathbf{\Phi}$	Comp Univ Score Var (%)
Plan	.8081	39.89%	.4650	37.61%
Act	.5108	9.40%	.0008	12.49%
Observe	.7802	26.10%	.7236	24.93%
Reflect	.7933	24.61%	.4010	24.97%
Composite	.8468	100.00%	.7823	100.00%

2.2 Decision Study (D-Study) Results

As shown in Table 5, the D-study results indicated that under the fully crossed design, three raters were sufficient to achieve an acceptable level of dependability for absolute decisions (**Φ** = .8057). Under the nested design, however, four raters per set were required to achieve a comparable level (**Φ** = .8067).

Table 5. Index of dependability (**Φ**) from the simulation of different numbers of raters.

$p^{\bullet} \times i^{\circ} \times r^{\bullet}$		$p^{\bullet} \times (i^{\circ} : r^{\bullet})$	
Number of raters	**Φ**	**Number of raters per set**	**Φ**
1	.5802	1 per set	.7308
2	.7343	2 per set	.7823
3	.8057	3 per set	.7899
4	.8468	4 per set	.8067

V. DISCUSSION AND CONCLUSION

The present findings indicate that the quality of classroom action research skill assessment, particularly the index of dependability and the variance structure of composite universe scores, depends substantially on measurement design. The fully crossed design ($p\bullet \times i\circ \times r\bullet$) yielded a higher index of dependability than the nested design ($p\bullet \times (i\circ : r\bullet)$) at both the composite and dimension levels. This pattern is consistent with the logic of Generalizability Theory and Multivariate Generalizability Theory, which suggests that design structure affects variance-component estimation, universe score variance, and the interpretation of score dependability for absolute decisions [22-23, 42-43]. In this sense, the study contributes to theory development in multidimensional performance assessment by showing that measurement design influences not only the magnitude of dependability but also the variance structure underlying composite universe scores.

The MFRM results also indicate that variation in rater severity and leniency was limited and that the raters functioned in a generally consistent manner prior to the multivariate generalizability analysis. This supports the logic adopted in the study that rater-mediated assessment should be examined diagnostically before score dependability is evaluated under MGT [30-31, 40]. At the same time, the fact that the fully crossed and nested designs still produced different levels of dependability suggests that score quality in this context was shaped not only by rater functioning itself but also by how raters were arranged across the multidimensional assessment structure. This interpretation is consistent with the argument developed in the Introduction and Theoretical Framework that rater-mediated performance assessment requires attention to both scoring quality and measurement design when scores are used for absolute decisions.

A further theoretical implication concerns the structure of the composite universe score. The results showed that the relative contribution of the PAOR dimensions differed across the two designs, indicating that the composite universe score was not simply the sum of independent dimension scores. Rather, the score structure reflected the multivariate relationships among dimensions under the specified measurement conditions. This point is important because it supports the MGT perspective that, even when dimensions are assigned equal nominal weight, their effective statistical contribution to the composite score may differ depending on the design [22-23]. Recent applications of MGT have likewise emphasized that multidimensional score interpretation should consider not only dependability coefficients but also the structural contribution of dimensions to composite scores [26-26, 35]. Accordingly, the present results suggest that the interpretation of multidimensional scores should consider both dependability and variance structure, particularly when the assessment is intended to support criterion-referenced interpretation and absolute decisions.

The study also extends the literature on measurement design optimization. In the present context, the fully crossed design provided stronger support for absolute decisions, whereas the nested design was less efficient in achieving comparable dependability and, in some dimensions, substantially altered the variance structure of the composite score. This suggests that optimization in multidimensional and rater-mediated assessment should not be based on the index of dependability alone. Instead, it should also consider the

intended use of scores, the arrangement of raters across dimensions, the preservation of dimension-level contribution to the composite score, and the practical constraints of implementation [22, 42-43]. From this perspective, the D-study results are particularly informative because they show that under the fully crossed design, at least three raters were sufficient to achieve a dependability level of approximately .80, whereas the nested design required a larger rater condition to reach a comparable level. These results reinforce the view that psychometric quality and practical feasibility must be considered together in design decisions.

A particularly important issue concerns the Act dimension. In both substantive and measurement terms, this dimension may be more difficult to assess from written classroom action research reports alone than the other PAOR dimensions. The assessment evidence in this study was derived from final written reports, whereas the Act dimension focuses on the implementation of the research plan in practice. Thus, the relatively low dependability observed for the Act dimension may reflect not only design-related influences but also a limitation in the alignment between the dimension being assessed and the form of evidence used for assessment. More specifically, the written reports may not have captured variation in actual implementation as clearly as they captured planning, observation, or reflection. In addition, the Act dimension may have shown more limited between-person variation in the present sample, which would reduce the proportion of variance attributable to student teachers and, in turn, lower the dependability estimate. Under the nested design, these constraints may have become even more pronounced because the design provided less favorable conditions for stable score interpretation in a dimension that was already more difficult to represent through written report evidence alone. In this respect, the present results suggest that dependability in multidimensional performance assessment may vary across dimensions partly because some dimensions are less fully represented in a single assessment format than others. They also indicate that score quality is shaped not only by facet structure and design but also by how well the selected evidence source represents each dimension of the construct [36-37].

However, this finding should be interpreted with caution. Because the present study examined only one nested configuration, it is not possible to determine conclusively whether the extremely low dependability of the Act dimension reflects a general limitation of nested designs in multidimensional performance assessment or, in part, an artifact of the specific dimension grouping used in this study. It is therefore more appropriate to interpret this result as evidence that nested arrangements may interact with dimension structure in important ways, rather than as evidence against all possible nested designs. Future research should compare alternative nested configurations, for example, Plan–Reflect with Act–Observe or Plan–Observe with Act–Reflect, to determine whether different dimension groupings yield more favorable variance structures and dependability estimates.

These results also have practical implications. When classroom action research skill assessment is used for individual-level absolute decisions, the fully crossed design appears more appropriate because it yields a higher index of dependability and better preserves the multidimensional score structure. However, when the purpose of assessment is developmental rather than high-stakes, lower levels of dependability may still be acceptable, particularly in contexts where fewer raters are available. In such cases, design decisions should be aligned with the intended use of scores and the level of decision impact [21-22, 24, 44]. This interpretation is consistent with the criterion-referenced emphasis of the study and with the broader principle that score quality should be judged in relation to the purpose and consequences of score use. For this reason, although the fully crossed design provided stronger support for individual-level absolute decisions, designs yielding lower dependability coefficients may still provide useful information in low-stakes contexts that emphasize feedback, reflection, and developmental use rather than formal classification.

These practical implications should be interpreted within the context of the present study. The results were based on 58 student teachers from one institution and one teacher education context, using four raters and one specific assessment instrument. Accordingly, the recommendation in favor of the fully crossed design should be viewed as provisional and context-bound rather than as a general prescription for all multidimensional performance assessments or teacher education settings.

Similarly, the D-study results indicate that the use of fewer raters may still be defensible when the assessment is intended to support learning and improvement rather than formal decision-making. Under the fully crossed design, one or two raters did not yield a level of dependability appropriate for individual-

level absolute decisions, yet such conditions may still provide useful information in classroom or program contexts where the main goal is to guide reflection on classroom action research performance. Under the nested design, lower-rated conditions may also be informative for developmental purposes, although the results suggest that this arrangement is less efficient when a stronger level of dependability is required. These results therefore support a more context-sensitive interpretation of design adequacy, in which the appropriate number of raters depends not only on psychometric criteria but also on the purpose and consequences of score use.

Nevertheless, several limitations should be acknowledged. First, the sample was drawn from a single institution, program, and specialization context, which limits the transferability of the findings across settings. Second, the assessment evidence was based on written classroom action research reports only. As noted above, this may have constrained the assessment of dimensions such as Act, which may require additional forms of evidence to be represented more adequately. Third, the study examined a limited range of rater conditions within the observed context. Future studies should therefore broaden the sample, compare alternative forms of assessment evidence for dimensions that are difficult to capture in written reports, and examine whether changes in score structure or weighting improve the dependability and interpretability of composite universe scores under alternative designs [25, 35, 43]. In particular, future research may benefit from combining written reports with additional evidence sources, such as implementation records, observational data, or mentor feedback, in order to better represent dimensions that are less fully captured in a written product alone.

In conclusion, the present study shows that measurement design has a substantial influence on classroom action research skill assessment, both in terms of the level of the index of dependability and the variance structure of composite universe scores. The results support the view that, in multidimensional performance assessment, design decisions affect not only how dependable scores are but also how the dimensions function together within the composite score [22-23, 25, 35, 43]. In particular, within the present assessment context, the fully crossed design, when using at least three raters, provided stronger support for individual-level absolute decisions than the nested design. At the same time, the results suggest that in low-stakes contexts, lower levels of dependability may still be acceptable when the primary goal is developmental use rather than formal classification. More broadly, the study indicates that measurement design optimization in multidimensional and rater-mediated assessment should consider psychometric quality, construct representation, evidence source, intended score use, and practical feasibility together rather than relying on a single coefficient alone.

Funding Statement

This research received no external funding. The APC was funded by the authors.

Author Contributions

Conceptualization, R.K. and S.K.; methodology, R.K., K.T., and S.K.; software, R.K.; validation, K.T. and S.K.; formal analysis, R.K.; investigation, R.K. and K.T.; resources, K.T. and S.K.; data curation, R.K.; writing—original draft preparation, R.K.; writing—review and editing, K.T. and S.K.; visualization, R.K.; supervision, S.K.; project administration, K.T. and S.K.; funding acquisition, S.K.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Data are available from the authors upon request.

Acknowledgments

Not applicable.

REFERENCES

1. Ahmad, H., & Guzman, S. T. (**2025**). Designing trustworthy educational artificial intelligence: A systemic framework for explainability, adaptivity, and ethical learning analytics. *Qubahan Techno Journal*, *4*(3).
2. McAteer, M. (**2013**). *Action research in education*. SAGE.
3. Wongwanich, S., Phiromsombat, C., & Srikhleub, K. (**2017**). *Development of a learning package to enhance classroom research skills of pre-service teachers*. Chulalongkorn University Press.
4. Cochran-Smith, M., & Lytle, S. L. (**2009**). *Inquiry as stance: Practitioner research for the next generation*. Teachers College Press.
5. Smit, B. H. J., Meirink, J. A., Tigelaar, D. E. H., Berry, A. K., & Admiraal, W. F. (**2024**). Principles for school student participation in pre-service teacher action research: a practice architecture's perspective. *Educational Action Research*, *32*(2), 222–242.
6. Lewin, K. (**1946**). Action research and minority problems. *Journal of Social Issues*, *2*(4), 34–46.
7. Kemmis, S., & McTaggart, R. (**1988**). *The action research planner* (3rd ed.). Deakin University.
8. Chanchusakun, S., & Varasunun, P. (**2020**). Performance Assessment: From Principle to Practice Guidelines. *Journal of Educational Measurement Mahasarakham University*, *26*(2), 36–56.
9. Thephasadin Na Ayudhya, W. (**2012**). *Classroom Action Research*. Dhurakij Pundit University Press.
10. Miller, D. M., Linn, R. L., & Gronlund, N. E. (**2009**). *Measurement and Assessment in Teaching* (10th ed.). Pearson.
11. Panadero, E., Jonsson, A., Pinedo, L., & Fernández-Castilla, B. (**2023**). Effects of rubrics on academic performance, self-regulated learning, and self-efficacy: A meta-analytic review. *Educational Psychology Review*, *35*, Article 113.
12. Uludag, P., & McDonough, K. (**2022**). Validating a rubric for assessing integrated writing in an EAP context. *Assessing Writing*, *52*, 100609.
13. Yılmaz, F.N. (**2024**). Comparing the reliability of performance task scores obtained from rating scale and analytic rubric using the generalizability theory. *Studies in Educational Evaluation*, *83*, Article 101413.
14. Li, W. (**2022**). Scoring rubric reliability and internal validity in rater-mediated EFL writing assessment: Insights from many-facet Rasch measurement. *Reading and Writing*, *35*, 2409–2431.
15. Huang, J., & Whipple, P. B. (**2023**). Rater variability and reliability of constructed response questions in New York state high-stakes tests of English language arts and mathematics: implications for educational assessment policy. *Humanities and Social Sciences Communications*, *10*, Article 860.
16. Mohd Noh, M. F., & Mohd Matore, M. E. E. (**2022**). Rater severity differences in English language as a second language speaking assessment based on rating experience, training experience, and teaching experience through many-faceted Rasch measurement analysis. *Frontiers in Psychology*, *13*, Article 941084.
17. Palermo, C. P. (**2022**). Rater characteristics, response content, and scoring contexts: Decomposing the determinates of scoring accuracy. *Frontiers in Psychology*, *13*, Article 937097.
18. Messick, S. (**1995**). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741-749.
19. Kane, M. T. (**2013**). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1-73.
20. Anderson, T. N., Lau, J. N., Shi, R., Sapp, R. W., Aalami, L. R., Lee, E. W., Tekian, A., & Park, Y. S. (**2022**). The utility of peers and trained raters in technical skill-based assessments: A generalizability theory study. *Journal of Surgical Education*, *79*(1), 206–215.
21. Nunnally, J. C., & Bernstein, I. H. (**1994**). *Psychometric theory* (3rd ed.). McGraw-Hill.
22. Brennan, R. L. (**2001**). *Generalizability theory*. Springer-Verlag.
23. Brennan, R. L. (**2006**). *Educational Measurement* (4th ed.). Praeger Publishers.
24. Shavelson, R. J., & Webb, N. M. (**1991**). *Generalizability theory: A primer*. SAGE.
25. Chen, D., Hebert, M., & Wilson, J. (**2022**). Examining human and automated ratings of elementary students' writing quality: A multivariate generalizability theory application. *American Educational Research Journal*, *59*(6), 1122–1156.
26. Kato, M. (**2022**). Examining the dependability and practicality of analytic rubric of summary writing using multivariate generalizability theory: Focusing on Japanese university students with lower-intermediate proficiency in English. *English Language Teaching*, *15*(9), 82–94.
27. Jiang, Z., Ouyang, J., Shi, D., Shi, D., Zhang, J., Xu, L., & Cai, F. (**2024**). Customizing Bayesian multivariate generalizability theory to mixed-format tests. *Behavior Research Methods*, *56*(7), 8080–8090.
28. Huang, H.-Y. (**2023**). Modeling rating order effects under item response theory models for rater-mediated assessments. *Applied Psychological Measurement*, *47*(4), 312–327.
29. Jiang, Z., & Skorupski, W. (**2018**). A Bayesian approach to estimating variance components within a multivariate generalizability theory framework. *Behavior Research Methods*, *50*, 2193–2214.

30. Anthony, C. J., Styck, K. M., Volpe, R. J., & Robert, C. R. (**2023**). Using many-facet rasch measurement and generalizability theory to explore rater effects for direct behavior rating–multi-item scales. *School Psychology*, *38*(2), 119–128.
31. Ramadhani, R., Syahputra, E., Simamora, E., & Soeharto, S. (**2023**). Expert judgement of collaborative cloud classroom quality and its criteria using the many-facets Rasch model. *Heliyon*, *9*(10), Article e20596.
32. Gordon, R. A., Peng, F., Curby, T. W., & Zinsser, K. M. (**2021**). An introduction to the many-facet Rasch model as a method to improve observational quality measures with an application to measuring the teaching of emotion skills. *Early Childhood Research Quarterly*, *55*, 149–164.
33. Brookhart, S. M. (**2015**). *Performance assessment: showing what students know and can do*. West Palm Beach: Learning Sciences International.
34. Hamzah, M. S. G., Idris, N., Abdullah, S. K., Abdullah, N., and Muhammad, M. M. (**2015**). Development of the double layer rubric for the study on the implementation of school-based assessment among teachers. *US-China education review*, *5*(4), 245–256.
35. Eskin, D. (**2022**). Generalizability of writing scores and language program placement decisions: Score dependability, task variability, and score profiles on an ESL placement test. *Studies in Applied Linguistics and TESOL*, *21*(2), 21–42.
36. Klyprayong, R. (**2026**). *Development of classroom action research skill assessment model of student teachers: An application of multivariate generalizability theory* (Unpublished doctoral dissertation). Chulalongkorn University.
37. Klyprayong, R., Tangdhanakanond, K., & Kanjanawasee, S. (**in press**). Development of the classroom action research skills assessment with classroom action research process components and the double layer scoring rubric. *Silpakorn Educational Research Journal*.
38. Jintanaprasert, P. (**2021**). *Self-assessment using different rubric methods on the development of mathematical problem-solving skills: Annotated and double-layer approaches* [Master's thesis, Chulalongkorn University]. Chulalongkorn University Intellectual Repository (CUIR).
39. Wancham, K., and Tangdhanakanond, K. (**2023**). Development of a two-tier rubric scoring criterion for assessing physics problem-solving ability. *Graduate Studies Journal, Valaya Alongkorn Rajabhat University under the Royal Patronage*, *17*(1), 16–31.
40. Linacre, J. M. (**1994**). *Many-Facet Rasch Measurement* (2nd ed.). MESA Press.
41. Khamboonruang, A. (**2023**). Detecting differential rater severity in a high-stakes EFL classroom writing assessment: A many-facets Rasch measurement approach. *PASAA*, *66*, 5–36.
42. Raymond, M. R., & Jiang, Z. (**2020**). Indices of subscore utility for individuals and subgroups based on multivariate generalizability theory. *Educational and Psychological Measurement*, *80*(1), 67–90.
43. Brennan, R. L., Kim, S. Y., & Lee, W.-C. (**2022**). Extended multivariate generalizability theory with complex design structures. *Educational and Psychological Measurement*, *82*(4), 617-642.
44. AERA, APA, & NCME. (**2014**). *Standards for educational and psychological testing*. American Educational Research Association.

Appendix. An example of a double-layer scoring rubric for classroom action research skill assessment.

Layer 1

PAOR	Items	Behavioral indicators	Level of score 3 Good	2 Satisfactory	1 Needs improvement
Plan	Study and analyze classroom problems	Examine and analyze classroom problems using supporting evidence	Clearly examines and analyzes classroom problems using evidence, provides supporting rationale, and identifies references	Examines and analyzes classroom problems using evidence and supporting rationale but does not clearly identify references	Examines and analyzes classroom problems without sufficient supporting evidence
Act	Use of instructional techniques	Implement the research plan as specified	Implements the research systematically according to the	Implements the research according to the specified plan but with some omissions	Unable to implement the research according to the specified plan or

PAOR	Items	Behavioral indicators	Level of score 3 Good	2 Satisfactory	1 Needs improvement
			specified plan, completely and within the designated timeframe	and/or incomplete within the designated timeframe	exceeds the designated timeframe
Observe	Data collection	Collect data systematically according to the plan	Collects data systematically according to the plan, using instruments with verified quality (e.g., Validity, Reliability, Difficulty (if applicable), and Discrimination (if applicable))	Collects data systematically according to the plan, but quality verification of instruments is incomplete in some aspects (e.g., Validity or Reliability or Difficulty (if applicable) or Discrimination (if applicable))	Unable to collect data systematically according to the plan, or instruments lack quality verification
Reflect	Conclusion and reflection	Summarize research findings and align with research objectives	Accurately summarizes research findings and fully aligns with the research objectives in a comprehensive and precise manner	Summarizes research findings correctly but lacks full alignment or completeness with respect to the research objectives	Summarizes findings inaccurately or does not align with the research objectives

Layer 2

Dimensions of the classroom action research process	Maximum score	Quality level of classroom action research skill (Score interpretation)		
		Good	Satisfactory	Needs improvement
Plan	36	≥ 35.27	22.17 – 35.26	< 22.17
Act	12	≥ 11.88	9.51 – 11.87	< 9.51
Observe	9	≥ 8.75	5.88 – 8.74	< 5.88
Reflect	12	≥ 10.75	7.38 – 10.74	< 7.38
PAOR	69	≥ 66.65	44.94 – 66.64	< 44.94