MEDICAL EDUCATION ONLINE
2025, VOL. 30, 2479962
https://doi.org/10.1080/10872981.2025.2479962

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Reliability and validity of simulation-based Electrocardiogram assessment rubrics for cardiac life support skills among medical students using generalizability theory

Sethapong Lertsakulbunlue [a], Kaophiphat Thammasoon[b] and Anupong Kantiwong [a]

[a]Department of Pharmacology, Phramongkutklao College of Medicine, Bangkok, Thailand; [b]Department of Student Affairs, Phramongkutklao College of Medicine, Bangkok, Thailand

ABSTRACT

Simulation-based learning (SBL) is effective for EKG interpretation training in the advanced cardiac life support (ACLS) context, enhancing motivation, confidence, and learning outcomes. However, research on the psychometrics of assessment rubrics for ACLS skills among pre-clinical students is limited. This study investigates the validity and reliability of assessment rubrics for ACLS skills, including EKG interpretation, scenario and pharmacological management, and teamwork. An SBL course that integrates basic EKG interpretation into ACLS Stations was conducted at Phramongkutklao College of Medicine, utilizing high-fidelity mannequins to simulate realistic scenarios, enrolling 96 medical students. The course consisted of five independent stations, and each student was assessed once by two raters using ten-item assessment rubrics. The rubrics included three domains: (1) EKG and ACLS algorithm skills, (2) management and mechanisms of action, and (3) affective domains. Validity evidence on the content was gathered, and construct validity was confirmed with confirmatory factor analysis (CFA). Inter-rater and internal consistency reliability were calculated. Generalizability theory was utilized to analyse the data. Three expert reviews yielded an item-objective congruence index of 0.67–1.00, with iterative validation through alpha and beta tests. The CFA demonstrated a good fit, but two questions with loading factors below 0.30 were removed, resulting in an eight-item assessment form. An inter-rater correlation of 0.70 ($p < 0.001$) and a Cronbach's alpha of 0.76 was demonstrated. To achieve a Phi-coefficient ≥0.80, three raters and at least 10 items are required in a p×i×r crossed design. With eight items, r:(p×i) nested design reliability was 0.69, 0.79, and 0.83 for one, two, and three raters, respectively. While a single rater with 10 items achieved a Phi-coefficient of 0.74. The rubrics for assessing ACLS skills among pre-clinical students demonstrated acceptable validity and reliability. A condensed eight-item rubric with acceptable reliability is proposed as a practical tool for optimizing assessment in future evaluations relevant to the pre-clinical context.

ARTICLE HISTORY
Received 14 November 2023
Revised 14 February 2025
Accepted 11 March 2025

KEYWORDS
Medical students; pre-clinical program; rubrics; EKG; ACLS; simulation; generalizability theory; assessment

Introduction

Simulation-based learning (SBL) is a key component of pre-clinical medical education due to its provision of a safe and controlled environment for learning and practicing clinical skills. Moreover, SBL enables medical students to broaden their experience and enhance their confidence and decision-making abilities [1]. SBL incorporation into medical curricula has the potential to enhance the quality of medical education and better prepare students for careers as medical practitioners in the future [2]. However, there are still concerns regarding the reliability and validity of simulation-based performance assessment scores for practising physicians, both individually and as team members.

Advanced cardiac life support (ACLS) is a critical skill for healthcare providers, particularly physicians, encompassing electrocardiogram (EKG) interpretation, pharmacological management, and effective teamwork. In the medical field, SBL using high-fidelity mannequins has been obtained to study ACLS [3]. While ACLS training has been shown to improve EKG interpretation skills, most medical students undergo this training in their late clinical years [4]. Moreover, SBL has demonstrated benefits not only in EKG learning but also in pharmacological therapy within the ACLS context [5,6]. Consequently, Phramongkutklao College of Medicine (PCM) has recently developed a pre-clinical course to enhance students' skills within the context of ACLS station simulation-based scenarios.

Recognizing the benefits of SBL in the ACLS context, PCM designed a specialized course incorporating high-fidelity mannequins to simulate realistic ACLS scenarios. These simulations encompassed history taking, heart and lung sound assessments, EKG monitoring, scenario management, and evaluation of pharmacological knowledge.

CONTACT Anupong Kantiwong anupongpcm31@gmail.com Department of Pharmacology, Phramongkutklao College of Medicine, Bangkok 10400, Thailand

The course enrolled third-year pre-clinical students, all of whom were required to learn EKG interpretation and pharmacological concepts related to cardiac life support. While these topics were previously taught through a traditional learning approach, paper-based methods often lack the realism necessary to fully capture the essence of ACLS. Although EKG interpretation and pharmacological knowledge are also assessed through multiple-choice and constructed-response questions, this SBL course was developed to assess students across cognitive, psychomotor, and affective domains, offering a more comprehensive and realistic approach. Furthermore, medical students at the PCM are admitted directly from high school and enrolled in a six-year curriculum. Incorporating SBL in their third year could facilitate a smoother transition into the clinical phase of their education. This approach also offers early clinical exposure to pre-clinical students, which has the potential to enhance their learning outcomes and increase their motivation [7].

Despite the increasing use of SBL, there are still concerns regarding the reliability and validity of simulation-based performance assessment scores [8]. Prior research has extensively studied computer simulations and standardized patients for assessment purposes [8–10]. However, reliable tools for assessing SBL with high-fidelity mannequins are limited, particularly for pre-clinical students. Thus, specific assessment tools for SBL in a realistic ACLS context to capture not only cognitive skills but also psychomotor and affective skills are needed.

Rubrics are valuable tools for assessment, as they define explicit performance criteria and expectations across multiple domains, including cognitive, affective, and psychomotor while ensuring consistency in grading [11]. Aligned with outcome-based education, rubrics are designed to reflect the competencies outlined in learning objectives [12]. The rubric development process in this study follows an analytical framework based on Association for Medical Education in Europe (AMEE) guidelines, ensuring detailed competency delineation across cognitive, psychomotor, and affective domains. Sub-competencies, such as EKG interpretation and communication skills, are aligned with the learning objectives [13], allowing each rubric criterion to be tailored to the expected level of expertise.

SBL assessment can assess knowledge, technical and clinical skills, communication, decision-making, patient safety, and teamwork skills, as well as higher-order competencies [14,15]. Assessment in SBL is widely utilized for formative and summative assessment [14], with summative assessment used for determining whether students have achieved the course's learning outcomes. However, utilizing SBL requires multiple raters and is resource-intensive, making it time-consuming [16]. Therefore, it is crucial to establish assessment tools' reliability, validity, and generalizability across varying numbers of raters and items.

Generalizability theory (G-theory) is a statistical method utilized to assess the reliability and validity of assessment tools in medical education. It is a theory that evaluates various assessment sources of variance, such as the rater, item, and student, and estimates how much of a contribution each makes to the overall variability in scores. Furthermore, a decision study (d-study) can help locate assessment error and inefficiency causes and forecast the best assessment layouts and scoring criteria [17]. It is beneficial for difficult tests, such as those involving several domains or skills like SBL [18]. G-theory has been extensively employed in medical education research to assess the quality of assessments and enhance their design and implementation.

Most G-theory studies focus on the number of occasions and raters as facets. However, at PCM, the high learner-to-teacher ratio and the limited number of Doctor of Medicine instructors pose unique challenges, resulting in only a single assessment occasion. Furthermore, SBL is an expensive and resource-intensive learning method [19]. Herein, the current study aims to determine the validity and reliability of the assessment rubrics design for assessing ACLS skills of pre-clinical medical students enrolled in the incorporating basic EKG interpretation into the ACLS station SBL course. Additionally, it aims to identify how we can generalize the medical students' overall scores in SBL across raters and assessment tool items. Thus, identifying the minimum number of raters and items would remarkably benefit resource management in SBL learning courses in the future.

Methods

Study design and subjects

The present study utilizes a G-theory analysis to determine the reliability of the assessment scores for incorporating basic EKG interpretation into the ACLS station course. A sample size of 43 was required for an effect size of 0.63 with 80% power at a significance level of 0.05 for a Pearson correlations test using G*Power 3.1.9.7 [20,21]. The SBL course was conducted during the cardiovascular system block at PCM, with 96 third-year pre-clinical students participating as part of their designated curriculum. The cardiovascular system block spanned four weeks. The first week focused on foundational basic science knowledge, including the anatomy and physiology of the cardiovascular system. The second and third weeks covered history taking, physical examination, pathology, and pharmacology related to the cardiovascular system. These weeks also incorporated team-based learning (TBL) and problem-

based learning (PBL) sessions centered on essential clinical cases. In the final week, project-based learning activities and the SBL class, along with examinations, were conducted. However, due to limited time and resources, students attended the SBL class only once during this block. The class aimed to achieve three objectives: (1) interpreting common EKGs within ACLS situations; (2) providing appropriate management based on the ACLS algorithm (verbally); and (3) understanding the mechanism of action of pharmacological treatments utilized in the ACLS algorithm and other common cardiac diseases, including ST-elevated myocardial infarction.

Preparations

The students were divided into ten groups, each consisting of two teams of four to five students. Prior to the class, students were provided with a briefing and assigned self-directed learning tasks on the relevant topics. They also completed e-learning assignments, which included exercises on interpreting common EKGs as outlined in the ACLS algorithm and studying the guidelines and steps for EKG interpretation within the algorithm. Additionally, a lecture on pharmacological management and EKG interpretation within the ACLS context was delivered before the SBL session.

Simulator and equipment

The simulations employed high-fidelity mannequins to facilitate basic physical examinations and create a realistic learning environment. Following the ACLS algorithm, the simulations covered themes such as tachyarrhythmias with and without pulse, bradyarrhythmias, asystole, and pregnancy-related arrhythmias. Life-size simulators, including the Laerdal SimMan 3 G for adult cardiac arrhythmias and the Laerdal SimMom for obstetric arrhythmias, were used. The simulated clinical scenarios were designed to replicate acute conditions, requiring both diagnosis and initiation of treatment within a 5-minute period.

The simulators were operated through software programmed to mimic acute medical conditions, with mannequins connected to monitors displaying continuous electrocardiographic tracings. Additional parameters, such as 12-lead EKG, blood pressure and oxygen saturation, were available upon the trainee's request. All equipment used during the simulation, including defibrillators, syringes, and endotracheal tubes, was authentic and sourced from the clinical environment, enabling participants to retrieve and use it in real time.

Simulation-based learning process

Five stations were set up, each presenting two consecutive scenarios. Each station had two teams of four to five students, and each team underwent testing in each scenario and the team leader's performance was evaluated under exam conditions, constituting 4% of the summative assessment. Each station lasted approximately 30 minutes, consisting of a 5-minute introduction and scenario briefing, 15 minutes of scenario engagement, and 10 minutes of debriefing. Each student was once assessed as a team leader and rotated through other roles within the ACLS team. Each group of students went through all stations, and all students took turns in the role of team leader. their performance was assessed solely when acting as the team leader within a different given scenario. All students were evaluated using the same evaluation form, with two raters stationed at each ACLS station (instructors from PCM). The raters included eight Doctors of Medicine and two healthcare professors. The study flow is presented in Figure 1.

Rubrics development and assessment

The performance of students during the course was assessed using scoring rubrics adapted from the AMEE guide framework and pertinent literature [13,22,23]. The rubric underwent content validation and rigorous scrutiny by all raters and was approved for use. Before the class, three professors performed content validation of the assessment form using the item objective congruence (IOC) method. The IOC revealed scores above 0.50 for all items, ranging from 0.67 to 1.00. Subsequently, the authors revised the form accordingly. An alpha test, comprising two internists, and a beta test, involving a group of five fourth-year medical students, were conducted to validate and gather feedback on the scenarios and assessment forms. Finally, all raters participated in a meeting to standardize the assessment process and comprehensively understand the study's objectives and assessment form. All raters unanimously agreed upon the amendments to the assessment form.

The assessment form included three domains: (1) EKG and ACLS algorithm skills, (2) management and mechanisms of action, and (3) affective domains. It comprised a total of 10 questions, each assigned a maximum score of 10 points. The following were the questions in the assessment form: (1) order of EKG interpretation, (2) accurate EKG interpretation, (3) EKG diagnosis, (4) ACLS algorithm order, (5) scenario management, (6) correct pharmacological treatment, (7) pharmacological mechanism of action, (8) interpersonal skills, (9) communication skills, and (10) learning responsibility. The complete assessment form is displayed in Figure 2.

Statistical analysis

Reliability analysis

The data analyses were carried out using IBM SPSS Statistics for Windows, Version 29.0 (Armonk, NY:

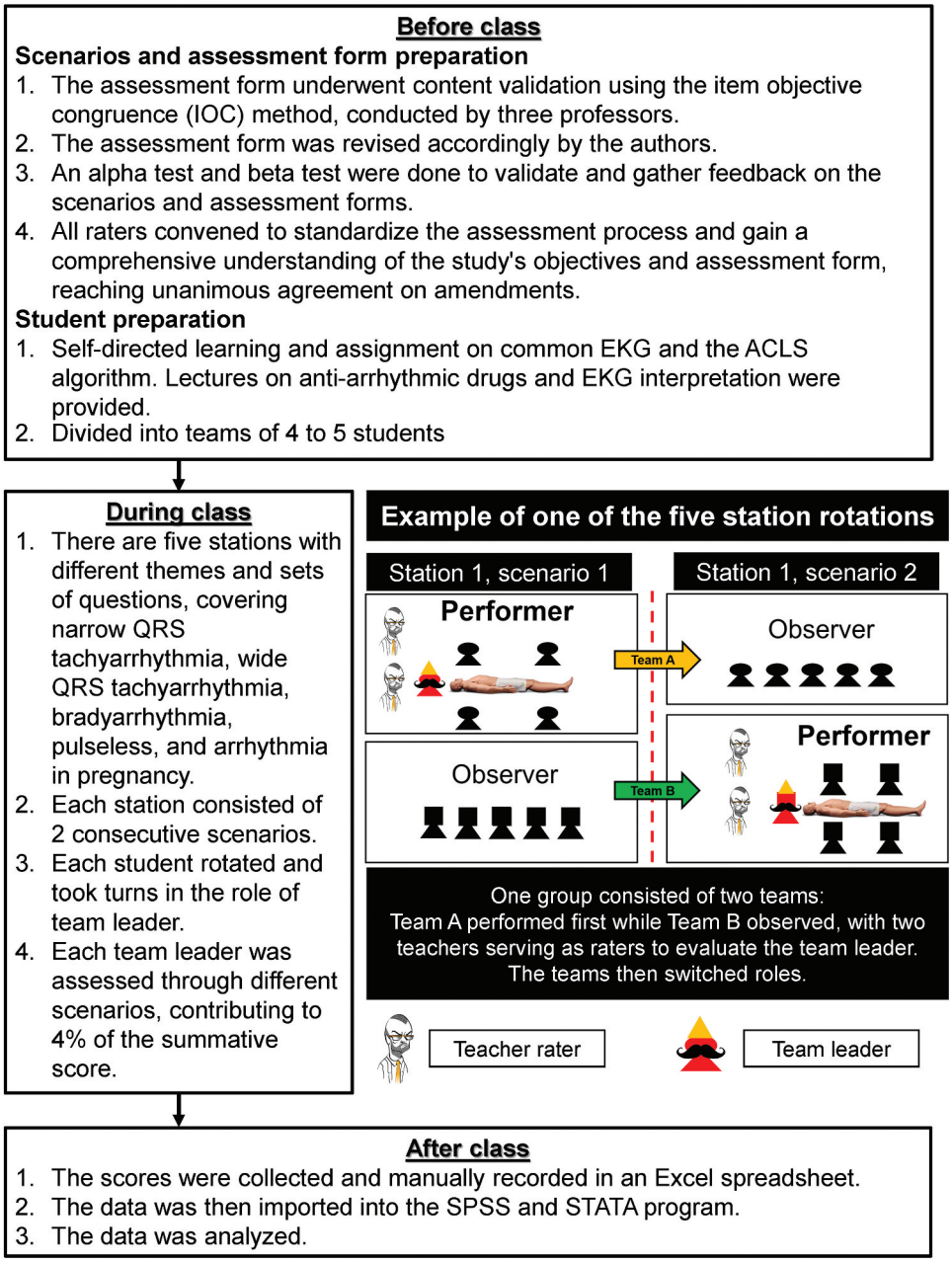

Figure 1. Study flow of integration of basic EKG interpretation into ACLS stations course.

IBM Corp). Categorical data were presented as percentages, and continuous variables were presented as means and standard deviations (SD). Inter-rater reliability was calculated using Pearson's correlation. Cronbach's alpha was utilized to identify the assessment instrument's internal reliability.

In order to test the reliability of the assessment instrument further, a G-theory analysis was carried out using a three-way ANOVA or p×i×r design, allowing for a fully crossed person (P) by questionnaire items (I) by raters (R) design. This analysis estimated the various aspects of measurement variance attributed to the study facets [24]. Estimate variance components were calculated [25]. In addition, a two-facet crossed design decision study was utilized for testing how the absolute G coefficient (Phi-coefficient) can change under various facet conditions and how to optimize the measurement. A Phi-coefficient above 0.80 indicates good reliability. Furthermore, a two-facet nested design (r:(p×i)) was employed [17]. Figure 3 reveals the sources of variability for the r:(p×i) design adapted from Brennan [17], where each person ($N = 96$) is evaluated by all items ($N = 10$). However, the subset of the raters (5 subsets of 2 raters) is different for each combination of persons and items. Person is crossed with items. Raters are nested within the combination of persons and items.

Validity analysis

CFA using maximum likelihood extraction was performed using *StataCorp, 2021, Stata Statistical*

Item	Score		
	10%	8-6%	4-0%
1. Order of EKG interpretation	Correct order of EKG interpretation	Some pattern of EKG interpretation is incorrect	Unable to complete EKG interpretation
2. EKG interpretation	Correct EKG interpretation	Incomplete EKG interpretation	Unable to interpret an EKG leading to incomplete diagnosis
3. EKG diagnosis	Correct diagnosis of abnormal conditions	Incomplete diagnosis of abnormal conditions	Unable to diagnosed the abnormal conditions
4. Order of ACLS algorithm	Adhered to ACLS algorithm	Some ACLS algorithm deviations	Incorrect ACLS practice
5. Scenario management	Manage the scenario correctly	Some incorrect management	Unable to manage the scenario
6. Pharmacologic treatment	Correct first-line pharmacological treatment	Incorrect pharmacological treatment of choice	Inappropriate pharmacological treatment
7. Mechanism of action	Able to explain the mechanism of action of the treatment and relate it to the scenario	Partial explanation of the mechanism of action of the treatment or cannot relate it to the scenario	Can't explain the mechanism of action of the treatment and relate it to the scenario
8. Interpersonal skills	Work efficiently with the members of the group and demonstrate good leadership/followership qualities in appropriate situations.	Works effectively with group members, demonstrates some leadership/followership qualities, communicates effectively and seeks help when needed	Struggles to work effectively with group members, lacks understanding of leadership/followership qualities, and fails to communicate effectively or seek help when needed
9. Communication skills	Exceptional communication skills, builds strong rapport with team members, conveys information clearly and effectively	Basic communication skills, establishes rapport with team members, conveys information effectively	Poor communication skills, fails to establish rapport with team members, struggles to convey information effectively
10. Responsibility	Thorough preparation, comprehensive understanding, takes leadership role and actively contributes to the course	Adequate preparation, basic understanding, takes responsibility and contributes to the course	Inadequate preparation, lacks understanding and little motivation to contribute to the course

Figure 2. English version assessment form used in integration of basic EKG interpretation into ACLS stations course.
I had uploaded a higher quality version of this figure. Please replace it for me.

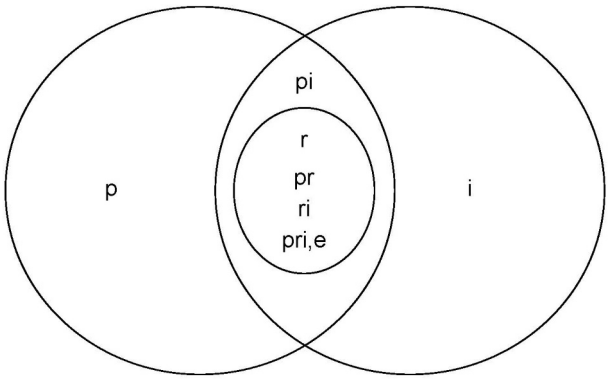

Figure 3. Sources of variability of nested r:(pxi) design. Each person ($N = 96$) is evaluated by all items ($N = 10$), but the subset of the raters (5 subsets of 2 raters) is different for each combination of persons and items. Person is crossed with items. Raters are nested within the combination of persons and items.

Software: Release 17 (College Station, TX: StataCorp LLC) to identify the construct validity. The current study employed six indices to assess the model fit, including (1) the chi-square test, represented as $\chi 2$; (2) the chi-square test over degree of freedom (df), indicated as $\chi 2/df$; (3) the comparative fit index, denoted as CFI; (4) the Tucker – Lewis index, also known as TLI; (5) the root-mean square error of approximation, or RMSEA; and (6) the root-mean square residual, referred to as SRMR. A good fit between the data and the hypothesized model was indicated by a $\chi 2/df$ value less than 2, a CFI value greater than 0.95, a TLI value greater than 0.95, an RMSEA value less than 0.05, and an SRMR value less than 0.05 [26,27].

Results

Reliability analysis

Inter-rater reliability and internal consistency reliability

A total of 192 ratings were completed by two raters assessing 96 students, with average scores of 9.53 ± 0.49 and 9.42 ± 0.47 for raters 1 and 2, respectively. The inter-rater correlations for each domain are presented in Table 1. An inter-rater correlation of 0.70 ($p < 0.001$) was observed, indicating a good correlation between the two raters.

For the internal consistency reliability, the overall Cronbach's alpha of the assessment form is 0.76, indicating acceptable internal consistency reliability. When assessing specific domains, Cronbach's alpha for the 'EKG and ACLS algorithm skills' domain was 0.70, which would increase to 0.78 if the 'order of ACLS algorithm' item was removed. Similarly, for the 'management and mechanism of action' domain, Cronbach's alpha was 0.62. However, removing the 'mechanism of action' item would result in Cronbach's alpha of 0.99. In the affective domain, Cronbach's alpha was 0.66, and all items had a total item correlation ranging from 0.44 to 0.50.

Table 1. Inter-rater reliability for the assessment of integration of basic EKG interpretation into the ACLS station by using simulation-based learning.

	rater 1	rater 2		
Domain	mean±SD	mean±SD	r	*p*-value
EKG and ACLS algorithm skills	9.36 ± 0.66	9.32 ± 0.64	0.492	<0.001
Scenario management and mechanism of action	9.61 ± 0.69	9.55 ± 0.58	0.458	<0.001
Affective domain	9.67 ± 0.60	9.41 ± 0.66	0.492	<0.001
Total	9.53 ± 0.49	9.42 ± 0.47	0.700	<0.001

SD: Standard deviation EKG: Electrocardiogram ACLS: Advanced cardiac life support.

Table 2. G-study for p × i × r and nested r:(p × i) design for the assessment form of integration of basic EKG interpretation into the ACLS station by using simulation-based learning, among 96 pre-clinical medical students, 10 items and 2 raters.

Source of Variation p×i×r design	*df*	SS	MS	Estimated Variance Component	% of Total Variance	Source of Variation r: (p×i) design	Estimated Variance Component	% of Total Variance
Student (P)	95	359.225	3.781	0.147	20.33	Student (P)	0.159	22.05
Item (I)	9	66.771	7.419	0.031	4.31	Item (I)	0.036	4.92
Rater (R)	1	5.002	5.002	0.004	0.50			
PI	855	513.129	0.600	0.092	12.69	PI	0.074	10.19
PR	95	62.498	0.658	0.024	3.34			
IR	9	11.290	1.254	0.009	1.21	Rater (R):PI	0.453	62.83
Residual (PIR, e)	855	356.210	0.417	0.417	57.62			
Total	1919	1374.125		0.723	100.00		0.721	100.00

SS Sum of squares, MS Mean of squares, df Degree of freedom

Generalizability study

Table 2 presents the findings of the two-facet G study for p×i×r and r:(p×i) designs for the overall score of the integration of basic EKG interpretation into the ACLS station using SBL. The findings reveal that the variance percentage attributable to the universe score, students (P), is 20.33% of the total variance. The effect of the variance component, including students (P) and the number of assessment tool items (I), is 12.69%. In the nested design, the variance components for students and items are 22.05% and 4.92%, respectively. The variance percentage of the interaction between students and items is 10.19%, whereas it is higher in the residuals (62.83%). These results suggest that the variation in the overall score is primarily due to the students themselves, with a relatively small contribution from the number of assessment items.

Table 3 presents the D study of p×i×r design, predicting the reliability of the instrument using various combinations of assessment items and raters. The table displays the phi-coefficient, ranging from 0.47 to 0.67 for one rater, 0.60 to 0.78 for two raters, and 0.66 to 0.82 for three raters, indicating that at least three raters are required to achieve a reliable assessment. However, Table 4 demonstrates the D study of r:(p×i) nested facet design, correlating with the course design of the present study. The results illustrate that a combination of 10 items and 2 raters is sufficient to achieve a reliable evaluation, with a Phi-coefficient of 0.83. While for three raters, only 8 items are needed (Phi-coefficient = 0.83). Figure 4 displays the G coefficient for the absolute decision for both p×i×r and r:(p×i) designs.

Validity analysis

Confirmatory factor analysis

The Kaiser-Meyer-Olkin test gave a value of 0.675, and the chi-square for Bartlett's test of sphericity

Table 3. D-Study of p×i×r design for assessment form of integration of basic EKG interpretation into the ACLS station by using simulation-based learning among pre-clinical medical students.

	Estimate variance Components In D-Study														
n_r'	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3
Effect n_i'	4	6	8	10	12	4	6	8	10	12	4	6	8	10	12
σ_p^2	0.147	0.147	0.147	0.147	0.147	0.147	0.147	0.147	0.147	0.147	0.147	0.147	0.147	0.147	0.147
σ_i^2	0.008	0.005	0.004	0.003	0.003	0.008	0.005	0.004	0.003	0.003	0.008	0.005	0.004	0.003	0.003
σ_r^2	0.004	0.004	0.004	0.004	0.004	0.002	0.002	0.002	0.002	0.002	0.001	0.001	0.001	0.001	0.001
σ_{pi}^2	0.023	0.015	0.011	0.009	0.008	0.023	0.015	0.011	0.009	0.008	0.023	0.015	0.011	0.009	0.008
σ_{pr}^2	0.024	0.024	0.024	0.024	0.024	0.012	0.012	0.012	0.012	0.012	0.008	0.008	0.008	0.008	0.008
σ_{ir}^2	0.002	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.000	0.000	0.001	0.000	0.000	0.000	0.000
σ_{pir}^2	0.104	0.069	0.052	0.042	0.035	0.052	0.035	0.026	0.021	0.017	0.035	0.023	0.017	0.014	0.012
$\hat{\sigma}^2_\delta$	0.151	0.109	0.088	0.075	0.066	0.087	0.062	0.050	0.042	0.037	0.066	0.046	0.037	0.031	0.027
$\hat{\sigma}^2_\Delta$	0.165	0.119	0.096	0.083	0.073	0.098	0.070	0.056	0.047	0.042	0.075	0.053	0.042	0.036	0.031
$E\rho^2$	0.493	0.575	0.626	0.662	0.689	0.628	0.703	0.748	0.777	0.799	0.691	0.760	0.799	**0.825**	**0.844**
Φ	0.471	0.552	0.604	0.640	0.667	0.601	0.678	0.725	0.756	0.778	0.661	0.734	0.776	**0.804**	**0.824**

n_r': Number of raters; n_i': Number of rubric items; $E\rho^2$: Relative G-coefficient; Φ: Absolute G-coefficient (Phi-coefficient).
Bold indicates acceptable reliability of 0.80 and above.

Table 4. D study of two-facet r:(p × i) nested design for the assessment form of the integration of basic EKG interpretation into the ACLS station using simulation-based learning among pre-clinical medical students.

Effect			Estimate variance Components In D-Study														
		n_r'	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3
p×i×r		r:(p × i)	n_i'	4	6	8	10	12	4	6	8	10	12	4	6	8	10	12
σ_p^2	0.153	σ_p^2	0.159	0.159	0.159	0.159	0.159	0.159	0.159	0.159	0.159	0.159	0.159	0.159	0.159	0.159	0.159
σ_i^2	0.035	σ_i^2	0.036	0.009	0.006	0.004	0.004	0.003	0.009	0.006	0.004	0.004	0.003	0.009	0.006	0.004	0.004	0.003
σ_r^2	0.005																	
σ_{pi}^2	0.088	σ_{pi}^2	0.074	0.018	0.012	0.009	0.007	0.006	0.018	0.012	0.009	0.007	0.006	0.018	0.012	0.009	0.007	0.006
σ_{pr}^2	0.023																	
σ_{ir}^2	0.007	$\sigma_{r:pi}^2$	0.453	0.113	0.076	0.057	0.045	0.038	0.057	0.038	0.028	0.023	0.019	0.038	0.025	0.019	0.015	0.013
σ_{pir}^2	0.471																	
$\hat{\sigma}_\delta^2$				0.132	0.088	0.066	0.053	0.044	0.075	0.050	0.038	0.030	0.025	0.056	0.037	0.028	0.022	0.019
$\hat{\sigma}_\Delta^2$				0.141	0.094	0.070	0.056	0.047	0.084	0.056	0.042	0.034	0.028	0.065	0.043	0.033	0.026	0.022
$E\rho^2$				0.547	0.644	0.707	0.751	0.784	0.679	0.761	**0.809**	**0.841**	**0.864**	0.739	**0.809**	**0.850**	**0.876**	**0.895**
Φ				0.531	0.629	0.694	0.739	0.772	0.655	0.740	0.791	**0.826**	**0.850**	0.710	0.786	**0.830**	**0.859**	**0.880**

n_r': Number of raters; n_i': Number of rubric items; $E\rho^2$: Relative G-coefficient; Φ: Absolute G-coefficient (Phi-coefficient).
Bold indicates acceptable reliability of 0.80 and above.

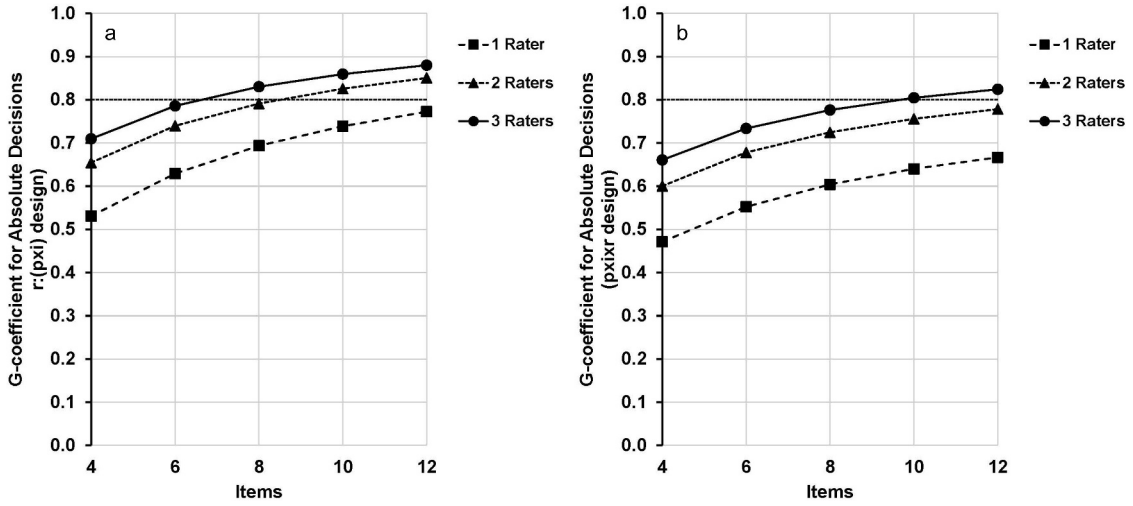

Figure 4. Decision study (D study) results for the pre-clinical medical students ($n = 96$) scores in integration of basic EKG interpretation into ACLS station course. Students were scored using two raters and ten assessment items. The coefficients are the projected phi-coefficient for different combinations of raters and assessment items. (a) r:(pxi) nested design (b) pxixr design.

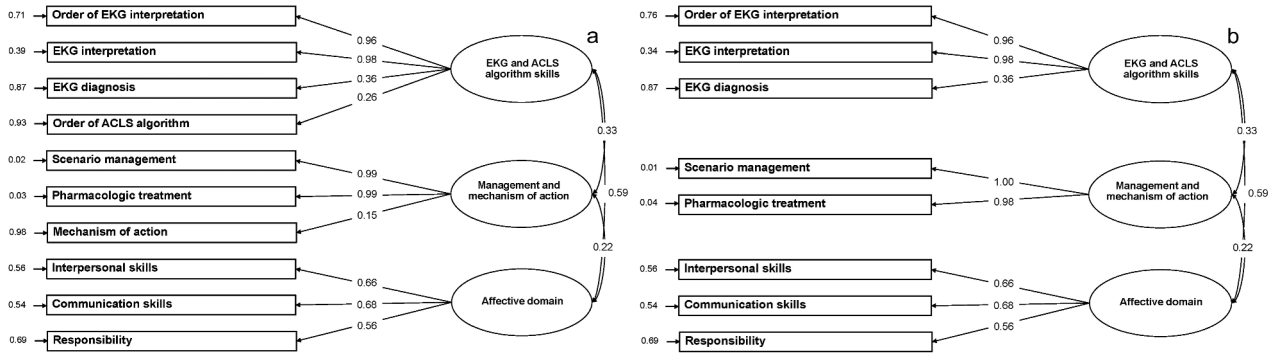

Figure 5. Confirmatory factor analysis of the full assessment form (a) and abbreviated assessment form (b) used in the integration of basic EKG interpretation into the ACLS station course. (a) The goodness of fit was tested, revealing the normed Chi-square value (χ^2/df) was 1.36, CFI = 0.99, TLI = 0.98, RMSEA = 0.05 and SRMR = 0.07, indicating an acceptable fit for the data. (b) The goodness of fit was tested, revealing the normed Chi-square value (χ^2/df) was 1.33, CFI = 1.00, TLI = 0.99, RMSEA = 0.04 and SRMR = 0.04, indicating a good fit for the data.

was significant ($\chi^2 = 1225.87$, df = 45, and $p < 0.001$), denoting a sufficiently large sample size for analysis ($N = 192$ ratings). The CFA of the complete assessment form was done and revealed acceptable goodness of fit, as displayed in Figure 5(a). Furthermore, in order to improve

the construct validity of the assessment form, questions with a loading factor below 0.3 were excluded. Consequently, a shortened eight-question assessment form was created with a goodness of fit index of normed chi-square value (χ^2/df) of 1.33, CFI = 1.00, TLI = 0.99, RMSEA = 0.04, and SRMR = 0.04, indicating a better fit for the data in comparison with the full 10-question form (Figure 5(b)).

Discussion

This study demonstrated the validity and reliability of assessment rubrics for evaluating students' overall performance in the 'Integration of Basic EKG Interpretation into ACLS Station' course. Both inter-rater and internal consistency reliabilities were acceptable, with values exceeding 0.70. Additionally, a generalizability study was conducted to determine the optimal number of items and raters required for reliable assessment, addressing constraints such as limited time and equipment in the SBL setting. The study revealed that a minimum of two raters are needed to attain acceptable reliability (Phi-coefficient ≥0.80). Moreover, the Phi-coefficient is relatively higher when evaluating only a subset of scenarios with two raters (a nested r:(p×i) design). Furthermore, this study shows that eight assessment items are sufficient to achieve good reliability with three raters. Thus, items with low loading factors were removed and left with an abbreviated form comprising eight questions. Using three raters on the abbreviated form resulted in good validity and reliability.

Reliability of the assessment rubrics

The overall inter-rater reliability in this study was 0.70, which is considered acceptable. However, the reliability within individual domains was comparatively lower, likely due to the challenges of assessing performance-based tasks as raters may struggle to provide consistent scores across different criteria [28]. To address this issue, it is crucial to develop detailed descriptors for each criterion in the rubric to enhance consistency and reliability. Providing good and poor performance examples for each criterion can enhance raters' understanding and improve evaluation accuracy and consistency. For instance, a previous SBL course to evaluate anesthesiology skills among residents employed a rater training program including calibration through independent video scoring until a consensus is reached, resulting in good inter-rater reliability across domains [29].

To our knowledge, this is the first study to apply G-theory analysis to evaluate ACLS simulation-based scenario scores among pre-clinical students, making direct comparisons with similar studies challenging. However, resource allocation challenges are not unique to this study. For instance, previous studies have utilized G-theory to optimize human resource allocation in pre-clinical Problem-Based Learning (PBL) programs and transthoracic echocardiography workplace-based assessments for physicians [30,31]. Additionally, G-theory analysis has been applied in SBL contexts where resources were extensively utilized [8,32,33]. While prior studies primarily focused on facets such as occasions or the number of scenarios, the current research addressed unique constraints, including a single assessment occasion per student due to a larger cohort and limited resources. To optimize these resources, this study focused on two key factors: the number of items and the number of raters.

Similar to other studies on performance-based assessments using scoring rubrics, the present study identified significant unexplained residual variance [8,30,32–34]. Moreover, the present study employed SBL in the ACLS context among pre-clinical students, which is relatively new for students and instructors, even with prior training. As a result, assessments, particularly in non-cognitive domains, may pose challenges, leading to slightly higher unexplained residual variance than SBL studies conducted with residents or physicians [32,33]. To mitigate this, these factors could be evaluated during debriefing rather than within scenarios to address the difficulty of assessing domains such as interpersonal skills, communication, and learning responsibility. Individual and peer-to-peer feedback sessions provide opportunities for assessing interpersonal and communication skills [35,36]. Additionally, refining the scale of assessment items could further enhance reliability, which will be explored in future studies.

The nested r:(p×i) design can exhibit higher reliability compared to a two-facet p×i×r crossed design in certain contexts. This is because having subsets of raters evaluate different students, rather than each rater assessing every student, can reduce potential rater fatigue and inconsistencies [24]. In the present SBL course, the design aligns with the r:(p×i) structure, where each rater evaluates specific scenarios for unique students. This alignment supports the expectation of higher reliability due to the focused assignment of raters to scenarios. Furthermore, in designing the SBL course, a nested design can help raters focus on a specific scenario, enhancing their expertise in assessing that scenario and potentially reducing the overall duration of the course. Therefore, a nested design should be considered when time or resource constraints are present.

Validity of the assessment rubrics

The rubric developed in this study is thoroughly aligned with the course learning objectives. Its

content validity was established through multiple approaches, including applying the AMEE guiding framework to inform rubric development and alignment with the Thai Medical Competency Assessment Criteria. Three experts conducted a content review using the Index of Item-Objective Congruence (IOC) method. Furthermore, alpha and beta testing were performed to ensure the rubric's reliability and robustness.

Regarding construct validity, a CFA was performed on the assessment form, which measured three different domains: EKG and ACLS algorithm skills, management and mechanism of action, and affective domain. Despite an acceptable fit of the overall model, two questions indicated low loading factors below 0.30. Nevertheless, these questions were retained in the G-theory model as they aligned with the learning objectives and literature review. However, in the nested design, we found that an eight-item assessment form with three raters achieved acceptable reliability. The present study suggested that employing an 8-item form could improve construct validity and obtain acceptable reliability, with the study's Phi-coefficient for the 8-item form with two raters being 0.79. Therefore, in practice, the questions with low loading factors could be removed to enhance construct validity and still achieve good reliability, resulting in an abbreviated eight-question form. Moreover, by reducing the number of items, the raters might be able to concentrate more on each item, thus reducing their burden [37].

Strategies to improve reliability

This study offers valuable insights into applying G-theory analysis for assessing pre-clinical SBL using high-fidelity mannequins. Limitations arose due to the high student-to-resource ratio, including mannequins, raters, and time, allowing only one assessment occasion per student. The findings indicate that achieving valid and reliable assessments requires either two raters for a ten-item rubric or three raters for an abbreviated eight-item version. However, the practicality of this approach is limited, as the simultaneous presentation of five scenarios would demand 10 to 15 raters. Consequently, strategies to enhance the reliability of the current rubrics are essential.

A previous study on SBL in anesthesiology suggested shortening scenarios to allow for the collection of additional performance samples and making scenario content more generic [8] to enhance reliability. However, since the current course is designed for pre-clinical students, who are relatively new to the ACLS context, this approach should be considered with caution. The present study integrated EKG interpretation with ACLS station mega codes, employing consecutive realistic scenarios to provide pre-clinical students with early clinical exposure. Alternatively, practice rounds or peer-led mock examinations could be introduced to enhance students' expertise and familiarity with the content before summative assessments. Including formative evaluations to assess students' achievement of learning outcomes has also been shown to improve satisfaction, as demonstrated in a study involving 218 undergraduate nursing students in clinical simulation learning [14].

Peer assessment through a multisource feedback process has been recognized as a reliable and valid method for evaluating the competencies of professionals and trainees [25]. In addition to instructor assessment, self- and peer assessment could be conducted. Moreover, it is crucial to give raters adequate training on how to utilize the assessment form and apply the criteria consistently. Conducting regular calibration exercises and providing feedback can help ensure that the raters consistently apply the criteria [38,39]. Additionally, conducting a formative assessment prior to the summative evaluation could benefit students. In the context of formative assessments, a Phi-coefficient of 0.70 is considered acceptable. The results of the present study indicate that only one rater is needed to achieve this level of reliability for the course's formative assessment.

Another approach to increase reliability is to provide a more detailed scoring system for each domain in the assessment form [39]. Previous research on the assessment of fourth-year medical students using a checklist exhibited good reliability in evaluating EKG interpretation skills [23]. The use of checklists has also demonstrated high reliability in objective structured clinical examination (OSCE) evaluation [39]. Moreover, detailed scoring checklists used during ACLS certification exams have been shown to have good validity and reliability [40]. Nevertheless, rubrics offer clear guidelines and expectations for student performance, promote grading and evaluation consistency, provide specific feedback on students' strengths and weaknesses, and can be utilized as a teaching tool to help students understand task or assignment expectations and develop self-assessment skills [11]. Therefore, integrating a checklist within each evaluated domain may improve the assessment's reliability and validity.

Limitations

The current study has some limitations that need to be acknowledged. Firstly, the study sample only included third-year pre-clinical students in a specific educational setting (i.e., PCM). Therefore, further research is required to determine the generalizability of our findings across various educational settings and multiple study years, as well as the clinical environment and different cultures. Secondly, the present study exclusively assessed students' EKG interpretation skills within the context of ACLS. Therefore, caution may be necessary when proceeding with the external validation of the study results.

Thirdly, the limited duration of the course restricted the time available for a comprehensive evaluation of the reliability and validity of the assessment scale. Future studies with extended timelines may provide more robust evidence for blueprinting, standard setting, consequences, quality control, and prediction of later performance. Finally, the nested model used in the Generalizability study was not capable of capturing the potential effect of other variables, including the number of scenarios, occasions and the order in which the students were tested, on the overall score. Moreover, the variation in scenario difficulty might have affected the results. In addition, while increasing the number of items is known to enhance reliability, this study required additional raters to achieve similar reliability levels due to constraints on item quantity. However, relying on more raters is less feasible and highlights a limitation of the study. Hence, given the current findings, caution should be exercised when using the assessment tool as a single-occasion high-stakes summative assessment.

Conclusion

The present study contributes to the evidence on the validity and reliability of the developed rubrics for assessing cardiac life support skills among pre-clinical students and investigates the optimal number of items and raters required for reliable assessment using generalizability theory. The findings suggest that employing two raters to evaluate single-occasion student performance is necessary to achieve good reliability. Assigning a unique set of raters to each student within a nested design may enhance reliability while also reducing the time required for assessments. To optimize resource utilization, enhancing the level of detail in the assessment form or incorporating peer assessment may be beneficial. Additionally, we propose a condensed assessment form with eight items, demonstrating good validity and acceptable reliability, as a feasible option for future evaluations.

Acknowledgments

This work would not have been possible without the active support of Phramongkutklao College of Medicine faculty members and its academic leaders, who are too numerous to name individually.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The author(s) reported there is no funding associated with the work featured in this article.

Data availability statement

The datasets used and/or analyzed during the current study are available by reasonable request from the author via Sethapong.ler@pcm.ac.th.

Ethical approval

The study was approved by the Medical Department Ethics Review Committee for Research in Human Subjects, Institutional Review Board, RTA (Approval no. S023q/66_Exp), according to the international guidelines including the Declaration of Helsinki, the Belmont Report, CIOMS Guidelines, and the International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use – Good Clinical Practice (ICH-GCP). Due to the use of secondary data, a waiver of documentation of informed consent was used.

Notes on contributors

SL reviewed the literature, designed the study, collected the data, data analysis and wrote the manuscript. KT collected the data, developed the methodology framework and developed the manuscript. AK reviewed the literature, designed the study, data analysis and wrote the first draft.

ORCID

Sethapong Lertsakulbunlue http://orcid.org/0000-0002-9349-2088
Anupong Kantiwong http://orcid.org/0000-0003-1353-3650

References

[1] Barry Issenberg S, Mcgaghie WC, Petrusa ER, et al. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. Med Teach. 2005;27(1):10–28. doi: 10.1080/01421590500046924
[2] McGaghie WC, Issenberg SB, Barsuk JH, et al. A critical review of simulation-based mastery learning with translational outcomes. Med Educ. 2014;48(4):375–385. doi: 10.1111/medu.12391
[3] Boysen-Osborn M, Anderson CL, Navarro R, et al. Flipping the advanced cardiac life support classroom with team-based learning: comparison of cognitive testing performance for medical students at the University of California, Irvine, United States. J Educ Eval Health Prof. 2016;13, Article 11. 11. doi: 10.3352/jeehp.2016.13.11
[4] Smith MW, Abarca Rondero D. Predicting electrocardiogram interpretation performance in advanced cardiovascular life support simulation: comparing knowledge tests and simulation performance among Mexican medical students. PeerJ. 2019;7:e6632. doi: 10.7717/peerj.6632
[5] Alamrani MH, Alammar KA, Alqahtani SS, et al. Comparing the effects of simulation-based and traditional teaching methods on the critical thinking abilities and self-confidence of nursing students. J Nurs Res. 2018;26(3):152–157. doi: 10.1097/jnr.0000000000000231
[6] Sigmon J, Davis P, Pari-An B, et al. 1053: Virtual simulation in advanced cardiac life support pharmacy education Crit Care Med. 2022;50(1):524–524. doi: 10.1097/01.ccm.0000810536.08259.ba

[7] Tayade M, Latti R. Effectiveness of early clinical exposure in medical education: settings and scientific theories – review. J Educ Health Promot. 2021;10(1):117. doi: 10.4103/jehp.jehp_988_20
[8] Sinz E, Banerjee A, Steadman R, et al. Reliability of simulation-based assessment for practicing physicians: performance is context-specific. BMC Med Educ. 2021;21(1), Article 207. doi: 10.1186/s12909-021-02617-8
[9] Guiton G, Hodgson CS, Delandshere G, et al. Communication skills in standardized-patient assessment of final-year medical students: a psychometric study. Adv Health Sci Educ. 2004;9(3):179–187. doi: 10.1023/B:AHSE.0000038174.87790.7b
[10] Swanson DB, Norcini JJ, Grosso LJ. Assessment of clinical competence: written and computer-based simulations. Assess Eval High Educ. 1987;12(3):220–246. doi: 10.1080/0260293870120307
[11] Panadero E, Jonsson A. The use of scoring rubrics for formative assessment purposes revisited: a review. Educ Res Rev. 2013;9:129–144. doi: 10.1016/j.edurev.2013.01.002
[12] Harden JRCMHDM, M R. AMEE Guide No. 14: outcome-based education: part 5-from competency to meta-competency: a model for the specification of learning outcomes. Med Teach. 1999;21(6):546–552. doi: 10.1080/01421599978951
[13] Pangaro L, ten Cate O. Frameworks for learner assessment in medicine: AMEE Guide No. 78. Med Teach. 2013;35(6):e1197–e1210. doi: 10.3109/0142159X.2013.788789
[14] Arrogante O, González-Romero GM, López-Torre EM, et al. Comparing formative and summative simulation-based assessment in undergraduate nursing students: nursing competency acquisition and clinical simulation satisfaction. BMC Nurs. 2021;20(1), Article 92. doi: 10.1186/s12912-021-00614-2
[15] Gordon CJ, Ryall T, Judd B. Simulation-based assessments in health professional education: a systematic review. J Multidiscip Healthc. 2016;9:69–82. doi: 10.2147/JMDH.S92695
[16] Savoldelli GL, Naik VN, Hamstra SJ, et al. [Les barriéres à l'utilisation de la formation basée sur simulateur]. Can J Anesth. 2005;52(9):944–950. doi: 10.1007/BF03022056
[17] Brennan RL. Generalizability theory. (New York): Springer-Verlag; 2001.
[18] Prion SK, Gilbert GE, Haerling KA. Generalizability theory: an introduction with application to simulation evaluation. Clin Simul Nurs. 2016;12(12):546–554. doi: 10.1016/j.ecns.2016.08.006
[19] Al-Elq A. Simulation-based medical teaching and learning. J Fam Community Med. 2010;17(1):35–40. doi: 10.4103/1319-1683.68787
[20] Faul F, Erdfelder E, Lang A-G, et al. G*power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods. 2007;39(2):175–191. doi: 10.3758/BF03193146
[21] Li H, Xiong Y, Zang X, et al. Peer assessment in the digital age: a meta-analysis comparing peer and teacher ratings. Assess Eval High Educ. 2016;41(2):245–264. doi: 10.1080/02602938.2014.999746
[22] Hogan S, Lundquist LM. The impact of problem-based learning on students' perceptions of preparedness for advanced pharmacy practice experiences. Am J Pharm Educ. 2006;70(4), Article. 82. doi: 10.5688/aj700482
[23] Raupach T, Hanneforth N, Anders S, et al. Impact of teaching and assessment format on electrocardiogram interpretation skills. Med Educ. 2010;44(7):731–740. doi: 10.1111/j.1365-2923.2010.03687.x
[24] Briesch AM, Swaminathan H, Welsh M, et al. Generalizability theory: a practical guide to study design, implementation, and interpretation. J Sch Psychol. 2014;52(1):13–35. doi: 10.1016/j.jsp.2013.11.008
[25] Donnon T, McIlwrick J, Woloschuk W. Investigating the reliability and validity of self and peer assessment to measure medical students' professional competencies. Creative Educ. 2013;4(06):23–28. doi: 10.4236/ce.2013.46A005
[26] Diamantopoulos A, Siguaw JA. Introducing LISREL. Sage Publications, Ltd, 2000. doi: 10.4135/9781849209359
[27] Schumacker RE, Lomax RG. A beginner's guide to structural equation modeling: Fourth edition 3rd. ed. Routledge; 2010. doi: 10.4324/9780203851319
[28] Mancar S ARSLAN, Gülleroğlu HD. Comparison of inter-rater reliability techniques in performance-based assessment. Int J Assess Tools Educ. 2022;9(2):515–533. doi: 10.21449/ijate.993805
[29] Blum RH, Muret-Wagstaff SL, Boulet JR, et al. Simulation-based assessment to reliably identify key resident performance attributes. Anesthesiology. 2018;128(4):821–831. doi: 10.1097/ALN.0000000000002091
[30] Guldbrand Nielsen D, Jensen SL, O'Neill L. Clinical assessment of transthoracic echocardiography skills: a generalizability study. BMC Med Educ. 2015;15(1):9. doi: 10.1186/s12909-015-0294-5
[31] Kassab SE, Du X, Toft E, et al. Measuring medical students' professional competencies in a problem-based curriculum: a reliability study. BMC Med Educ. 2019;19(1), Article 155. doi: 10.1186/s12909-019-1594-y
[32] Andersen SAW, Park YS, Sørensen MS, et al. Reliable assessment of surgical technical skills is dependent on context: an exploration of different variables using generalizability theory. Academic Med. 2020;95(12):1929–1936. doi: 10.1097/ACM.0000000000003550
[33] Boulet JR, Murray D, Kras J, et al. Reliability and validity of a simulation-based acute care skills assessment for medical students and residents. Anesthesiology. 2003;99(6):1270–1280. doi: 10.1097/00000542-200312000-00007
[34] Murray DJ, Boulet JR, Avidan M, et al. Performance of residents and anesthesiologists in a simulation-based Skill assessment. Anesthesiology. 2007;107(5):705–713. doi: 10.1097/01.anes.0000286926.01083.9d
[35] Kilminster S, Cottrell D, Grant J, et al. AMEE Guide No. 27: effective educational and clinical supervision. Med Teach. 2007;29(1):2–19. doi: 10.1080/01421590701210907
[36] Uhm S, Lee GH, Jin JK, et al. Impact of tailored feedback in assessment of communication skills for medical students. Med Educ Online. 2015;20(1), Article 28453. 28453. doi: 10.3402/meo.v20.28453
[37] Cook KF, Kallen MA, Amtmann D. Having a fit: impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. Qual Life Res. 2009;18(4):447–460. doi: 10.1007/s11136-009-9464-4
[38] Schuwirth LWT, Van der Vleuten CPM. Programmatic assessment: from assessment of learning to assessment for learning. Med Teach. 2011;33(6):478–485. doi: 10.3109/0142159X.2011.565828
[39] Shumway JM, Harden RM. AMEE Guide No. 25: the assessment of learning outcomes for the competent and reflective physician. Med Teach. 2003;25(6):569–584. doi: 10.1080/0142159032000151907
[40] McEvoy MD, Smalley JC, Nietert PJ, et al. Validation of a detailed scoring checklist for use during advanced cardiac life support certification. Simul Healthc. 2012;7(4):222–235. doi: 10.1097/SIH.0b013e3182590b07