ACTA MÉDICA PORTUGUESA
A Revista Científica da Ordem dos Médicos

Development and Psychometric Validation of the Minho Suture Assessment Scale (Minho-SAS) for Medical Students

Desenvolvimento e Validação Psicométrica da Escala de Avaliação de Sutura do Minho (Minho-SAS) para Estudantes de Medicina

Nuno SILVA GONÇALVES [1,2,3], Rita MATOS SOUSA [1,2,3], Carlos COLLARES [4,5,6], Vítor Hugo PEREIRA [1,2,3], José Miguel PÊGO [1,2,5], M. BROWNELL ANDERSON [1]
Acta Med Port 2026 Jan;39(1):23-32 ▪ https://doi.org/10.20344/amp.23567

ABSTRACT

Introduction: Even though mastery of suturing is a core technical skill in surgical education, existing tools for its assessment often lack psychometric validation or are not specifically designed for undergraduate training. The aim of this study was to develop and validate the Minho Suture Assessment Scale (Minho-SAS), a structured instrument to evaluate fundamental suturing competencies in medical students. The research question was whether the Minho-SAS demonstrates validity and reliability as a psychometric tool.

Methods: The development process involved collaboration with multidisciplinary surgical teams and experienced practitioners to ensure content validity. Data from a cohort of medical students were utilized for psychometric evaluation. Dimensionality was assessed using parallel analysis, Bayesian information criterion, unidimensional congruence, item unidimensional congruence, explained common variance, item explained common variance and mean of item residual absolute loadings. Validity based on internal structure was assessed with Rasch model analysis and factor analysis from the tetrachoric correlation matrix. Reliability was assessed using Rasch model standard errors of measurement to obtain a conditional reliability curve and Cronbach's alpha and McDonald's omega internal consistency coefficients.

Results: Analyses supported a unidimensional structure for the Minho-SAS. The single-factor solution explained 39.96% of variance, and Rasch measures accounted for 29.15% (16.43% by persons, 12.72% by items). Residual correlations, factor loadings, and item fit statistics were within acceptable ranges. Reliability indices were satisfactory: Rasch reliability = 0.706; McDonald's omega = 0.889; Cronbach's alpha = 0.883.

Conclusion: The Minho-SAS is a robust instrument specifically tailored for assessing fundamental suturing skills among medical students. Rasch model analysis yielded less favorable results than factor analysis, yet still acceptable. While demonstrating considerable potential, further exploration of Minho-SAS across diverse populations and educational settings is crucial to affirm its broader applicability and impact in medical education and clinical practice.

Keywords: Clinical Competence; Educational Measurement; Students, Medical; Surveys and Questionnaires; Suture Techniques/education

RESUMO

Introdução: Apesar do domínio da sutura constituir uma competência essencial na educação cirúrgica, os instrumentos existentes para a sua avaliação carecem de validação psicométrica ou especificidade para ensino pré-graduado. Este estudo teve como objetivo desenvolver e validar a Escala de Avaliação de Sutura Minho (Minho-SAS), para avaliar competências de sutura em estudantes de Medicina, avaliando a sua validade e fiabilidade enquanto ferramenta psicométrica.

Métodos: O processo de desenvolvimento envolveu a colaboração com equipas multidisciplinares e experientes para assegurar a validade de conteúdo. Foram utilizados dados de um grupo de estudantes de Medicina para a avaliação psicométrica. A dimensionalidade foi analisada através de análise paralela, critério de informação de Bayes, *unidimensional congruence*, *item unidimensional congruence*, variância comum explicada, variância comum explicada por item, e média dos resíduos absolutos dos itens. A validade com base na estrutura interna foi avaliada com análise do modelo Rasch e análise fatorial a partir da matriz de correlação tetracórica. A confiabilidade foi avaliada através dos erros padrão de medição do modelo Rasch para obtenção da curva de confiabilidade condicional e dos coeficientes de consistência interna alfa de Cronbach e ómega de McDonald.

Resultados: As análises sustentaram uma estrutura unidimensional. A solução unifatorial explicou 39,96% da variância, e as medidas de Rasch representaram 29,15% (16,43% - participantes e 12,72% - itens). As correlações residuais, cargas fatoriais e estatísticas de ajuste dos itens encontravam-se dentro de valores aceitáveis. Os índices de confiabilidade foram satisfatórios: Rasch = 0,706; ómega de McDonald = 0,889; alfa de Cronbach = 0,883.

Conclusão: A Minho-SAS revela-se um instrumento robusto, especificamente desenvolvido para avaliar competências de sutura em estudantes de Medicina. A análise Rasch apresentou resultados menos favoráveis que a análise fatorial, embora aceitáveis. Apesar do seu potencial considerável, é fundamental explorar a Minho-SAS em diferentes populações para confirmar a sua aplicabilidade e impacto mais amplos na educação médica e na prática clínica.

Palavras-chave: Avaliação Educacional; Competência Clínica; Estudantes de Medicina; Inquéritos e Questionários; Técnicas de Sutura/educação

1. School of Medicine. Universidade do Minho. Portugal.
2. Life and Health Sciences Research Institute (ICVS). School of Medicine. Universidade do Minho. Portugal.
3. Clinical Academic Center (2CA). Braga. Portugal.
4. Medical Education Unit. Faculty of Medicine and Biomedical Sciences. Universidade do Algarve. Faro. Portugal.
5. European Board of Medical Assessors. Cardiff. United Kingdom.
6. Inspirali Educação. São Paulo. Brazil.

✉ **Autor correspondente:** Nuno Silva Gonçalves. nunogoncalves@med.uminho.pt

Revisto por/*Reviewed by*: Miguel Castelo-Branco, Manuel Rosete
Recebido/*Received*: 13/07/2025 - **Aceite/*Accepted*:** 08/10/2025 - **Publicado *Online*/*Published Online*:** 18/12/2025 - **Publicado/*Published*:** 02/01/2026
Copyright © Ordem dos Médicos 2026

KEY MESSAGES

- Provides a standardized, detailed instrument specifically designed for evaluating basic suturing competence in medical students.
- Developed through a rigorous process involving multidisciplinary surgical experts and modern psychometric approaches.
- Combines classical test theory and item response theory analyses to ensure robust validity and reliability evidence.
- Aims to identify students who have not yet reached essential competence, supporting safer and more targeted surgical education.

INTRODUCTION

Suturing skills are an essential competence for every young doctor who is expected to be capable of autonomously treating a patient with a regular skin wound. For this reason, this skill must be developed and assessed during undergraduate medical training to ensure that medical students are prepared for this practice immediately after graduation. However, despite most medical educators recognizing this as an essential skill, little is known about how medical schools are assessing suture skills, leading to a problem in acquiring essential competencies. Despite the availability of various formats for assessing technical and surgical competencies in the literature, these tools are targeted towards more complex procedures, mostly applicable to surgery residents and not designed for undergraduates. To the best of our knowledge a suturing skill assessment scale targeted specifically at medical students is lacking. To address this gap, we propose to develop a new scale and obtain validity evidence based on its internal structure.

The proper assessment of competence is a challenging task.[1] This can be caused in part by the struggle to define and isolate the competency to be evaluated, since there are several dimensions for professional competency and even a hierarchy amongst them.[2,3]

In the past decades, efforts have been made to progress to competency-based curricula, such as the creation of CanMEDS,[4] Accreditation Council for Graduate Medical Education core competencies,[5] or Entrustable Professional Activities.[6] Epstein *et al* explained that competence is contextual, reflecting the relationship between a person's abilities and the tasks they are required to perform in a particular situation in the real world.[7] Regarding undergraduate performance assessment, students can be assessed in a simulated environment or a real-life one, and, for both, multiple instruments can be used to assess the students, depending on the competence being measured.

Regarding the surgical undergraduate curriculum, surgical skills are unquestionably one of the core components and one of the most relevant abilities expected and required for a medical graduate. Amongst these surgical skills, the ability to perform an acceptable basic suture that can be autonomously replicated in any scenario should be a requirement for practice. However, recent reports show a concerning trend in the decline of suturing competence among recent medical graduates.[8]

Several suturing assessment scales are available in the literature,[9] mostly directed to laparoscopic/robotic sutures,[10] microsurgical,[11-13] post-graduate surgical residents,[12,14-17] or even task-oriented.[18-20] Others address basic suturing skills and are targeted at medical students but include the time to perform as one of the components of competence.[21] The factor of 'time to perform' has been very controversial in the literature since the authors believe the expected competency for the undergraduate should be focused on the quality of the technique, not its speed.

Since competence is contextual and developmental, we consider that the assessment must be tailored to the context. However, when the assessed competence involves a specific technique, most assessment methods are designed with expert-level performance in mind—or at least with expectations of high proficiency—without accounting for varying stages of expertise development.[22] This has led us to believe that there is still space to discuss what can be done when the assessment purpose is to discriminate between those who are competent, disregarding the level of expertise, from those who are unable to perform. Additionally, we aim to improve the specificity, detail, and standardization of the assessment tool, reducing subjectivity and making it feasible for less experienced assessors to be used in summative exams, reducing the burden and logistic problems for the faculty and the medical school.

To properly validate and understand the measurement process in educational assessment, complementary approaches from classical test theory (CTT) and item response theory (IRT) provide valuable insights into how well an assessment tool performs.[23] While CTT offers a straightforward way to evaluate test reliability, modern psychometric approaches like the Rasch model analysis, a special case under the umbrella of IRT, can provide more sophisticated information about both the assessment items and student performance.[23]

In the context of surgical skills assessment, having precise and reliable measurement tools is crucial, particularly when making high-stakes decisions about student competency. The ability to accurately distinguish between competent and non-competent performance levels requires assessment tools that have undergone rigorous validation.[24] This validation should examine not only the basic reliability of the tool but also provide evidence that it measures the intended construct – in this case, basic suturing skills – in a consistent and meaningful way.[25] The internal structure of the assessment tool was examined in this study through multiple complementary dimensionality analyses, ensuring that the scale measures a single, coherent construct of basic suturing skills.

This work introduces the Minho Suture Assessment Scale (Minho-SAS), a new evaluation tool using binary (yes/no) items to assess whether undergraduate medical students are competent in basic suturing skills. The primary objective was to provide evidence of validity based on the internal structure of the scale, through both classical test theory and item response theory approaches. Secondary objectives included examining the reliability of the instrument, assessing its dimensionality, and evaluating its feasibility for use in summative and formative assessments within medical education. To our knowledge, this represents the first application of such a thorough psychometric framework to surgical skills assessment.

METHODS

Step 1: Identification of the purpose and domains of the assessment

We intended to create an assessment instrument, in the European Portuguese language, that could identify undergraduates able to perform a correct basic suture. The instrument should be user-friendly and adaptable to the context: summative assessment for 5th-year medical students in a simulated scenario. To the best of our knowledge, no other scale with this purpose and context has been validated.

For guidance on the best practices for scale validation, the work from Boateng *et al* and the standards for educational and psychological testing were used.[26,27]

The study was approved by the Ethics Committee for Research in Life and Health Sciences (CEICVS) (CEICVS 103/2022).

Step 2: Development of the items of the Minho Suture Assessment Scale

We used both inductive and deductive approaches simultaneously to generate the items. Fourteen skilled surgeons from various backgrounds, such as general surgery, pediatric surgery, gynecology, orthopedics, and urology, were carefully selected to form a focus group. The group employed the Delphi methodology to create a preliminary list of items for the Minho-SAS prototype. Each expert provided initial suggestions and feedback, then refined their views based on group discussions. In the first round, items

Table 1 – Minho Suture Assessment scale translated into the English language

Question	Item	Yes/No	No. of positive assessments
1	Confirms and prepares all necessary materials.	Y/N	242
2	Holds the needle correctly with the needle holder.	Y/N	251
3	Manipulates the needle holder correctly.	Y/N	259
4	Manipulates the tissue forceps correctly.	Y/N	245
5	Places the suture at an appropriate distance from the wound edge and ensures equal distance on both sides.	Y/N	221
6	Demonstrates surgical dexterity in inserting, exteriorizing, and handling the needle at the skin's exit.	Y/N	181
7	Ties the knot, counter-knot, and additional knots correctly.	Y/N	249
8	Cuts the thread to the proper length using correct technique.	Y/N	206
9	Positions the knot lateral to the wound.	Y/N	199
10	Applies the correct degree of tension to the knot.	Y/N	209
11	Maintains appropriate distance between sutures.	Y/N	209
12	Performs suturing safely throughout the procedure.	Y/N	236
13	Safely stores suturing equipment and needle upon completion of the procedure.	Y/N	166
14	Executes the suture with dexterity, economy, and fluidity of movement.	Y/N	164
15	Demonstrates respect for tissues throughout the procedure.	Y/N	228
Global	Global Assessment Score	1 to 5	-

were introduced for discussion. Duplicates were removed, wording was standardized, and a consensus was reached on each item. The Delphi process consisted of three rounds, during which experts reviewed and discussed the items, offering feedback and assessing relevance. By the third round, consensus was achieved, and the final draft was collectively revised and approved by the group.

The final draft contained 15 yes/no items and a global performance scale with a Likert-type rating ranging from 1 to 5 (Table 1).

Step 3: Sample Size justification and application of the scale

We followed a 10:1 rule of thumb for determining the sample size,[28] which implied a minimum of 150 assessments, and used the Minho-SAS to evaluate the performance of 269 fifth-year medical students who underwent an objective structured clinical examination (OSCE) during a surgical clerkship. The assessments were carried out by six experienced medical doctors who were trained in the technique and had extensive experience in assessment.

Step 4: Psychometric analysis

The dimensionality of the assessment tool was evaluated using a comprehensive set of complementary psychometric methods. We began with parallel analysis, which helps determine the optimal number of dimensions by comparing the eigenvalues of our data against those from randomly generated datasets. This analysis was supplemented by the Bayesian information criterion (BIC), which helped evaluate different dimensional models by balancing model fit and complexity. To gain deeper insights into the unidimensional structure, we employed several specialized indices: unidimensional congruence (UniCo) and item unidimensional congruence (I-UniCo), which assess how well the data fit a single-dimension model at both the overall test and individual item levels; explained common variance (ECV) and item explained common variance (I-ECV), which quantify how much of the variance can be attributed to a primary dimension; and mean of item residual absolute loadings (MIREAL), which examines whether significant secondary dimensions exist. This multi-method approach provides robust evidence for the internal structure validity of the assessment tool, giving us confidence in our understanding of what the scale is measuring.

To obtain validity evidence based on the internal structure of the scale, we first conducted an item response theory analysis that included Rasch and two-parameter logistic (2-PL) models. Item parameters (difficulty and discrimination), item characteristic curves, and item information functions were estimated. Item fit was evaluated using infit and outfit statistics, with values between 0.5 and 1.5 considered acceptable. The conditional reliability curve was generated to examine measurement precision across the ability spectrum.

Factor analysis of the tetrachoric correlation matrix was also performed, due to the dichotomic nature of the items. Due to the presence of asymmetric distributions in the items (as evidenced by skewness values ranging from -4.911 to -0.451), as well as extreme values for kurtosis (ranging from -1.793 to 22.032), a robust estimation method, robust diagonally weighted least squares (RDWLS), was employed for factor extraction. This method used robust mean and variance-scaled corrections for chi-square calculations. Bootstrapping with 1000 samples was used to obtain bias-corrected confidence intervals for key parameters. The rotation method was Robust Promin, using Weighted Varimax as the clever rotation start.

In addition to the validity based on internal structure, we aimed to obtain additional validity evidence based on:

A) Test content: To assure content validity, adding to the approval of the previously described focus group, the scale was approved by additional expert medical professionals with experience in suture and psychometrics, ensuring scale alignment with the content.
B) Response processes: To ensure the fit between the assessment process and the construct, assessors were subsequently questioned about their thought process, if they were evaluating suture skills, and if they thought they were assessing other competencies. Additionally, assessors were accompanied by surgical experts during the first assessments.
C) Relations with other variables: To ensure the accuracy of the results, we used an established and validated technical assessment scale called objective structured assessment of technical skill (OSATS) in combination with the Minho-SAS.[29] This approach provided additional validity evidence to Minho-SAS and helped to enhance the validity of the findings. We conducted a correlation analysis and calculated Cohen's d between the test scores of the two assessed years to further validate the results.
D) Consequences of testing: Following the analysis of the test scores, the faculty members responsible for the curricular unit blueprint and mapping can modify the curriculum to address possible gaps in students learning, as intended consequences of this exam. In the future, the scale can also be tested for feedback delivery.

We used Cronbach's alpha and McDonald's omega internal consistency coefficients and Rasch model scores' standard errors of measurement to estimate the reliability of the scale, in which coefficients above 0.70 were regarded as acceptable.[27]

The analyses were performed using multiple software packages: Winsteps was used for the principal component analysis of Rasch model residuals (PCAR) and item-person maps. FACTOR version 12.02.01 was used for factor analysis of the tetrachoric correlation matrix and other dimensionality analyses and the R package *mirt* (R version 4.3.1) for the remaining Rasch and 2-PL model analyses. The R script used can be found in Appendix 1 (Appendix 1: https://www.actamedicaportuguesa.com/revista/index.php/amp/article/view/23567/15832). Classical reliability analyses using McDonald's omega and Cronbach's alpha coefficients and their confidence intervals were conducted using JASP (Version 0.18.1). Correlations between scores and Cohen's *d* coefficients were obtained using jamovi (v2.3.28).

RESULTS

The final Minho-SAS, composed of 15 dichotomous yes/no items and a global performance score ranging from 1 to 5, is presented in Table 1, with additional information regarding the number of positive assessments in the OSCE. The Minho-SAS is also available in Appendix 2 (Appendix 2: https://www.actamedicaportuguesa.com/revista/index.php/amp/article/view/23567/15833).

For the initial validation, the face-validation process was performed by showing the final prototype to experienced surgeons, with expertise on the skill to be assessed, with an unanimous positive judgement.

Dimensionality analysis

Factor analysis of the tetrachoric correlation matrix supported a unidimensional structure. The single-factor solution explained 39.96% of the total variance (eigenvalue = 5.99). Unidimensionality indices met established criteria: UniCo = 0.967 (95% CI: 0.968 - 0.979), ECV = 0.871 (95% CI: 0.875 - 0.884), and MIREAL = 0.184 (95% CI: 0.140 - 0.204). Model fit indices were adequate: RMSEA = 0.039 (90% CI: 0.000 - 0.058), CFI = 0.943 (95% CI: 0.875 - 1.190), and GFI = 0.911 (95% CI: 0.882 - 0.959). Factor loadings ranged from 0.213 to 0.822.

Q3 residual correlations from Rasch analysis were all below 0.3, supporting local independence. The principal component analysis of Rasch model residuals revealed an eigenvalue of 1.59 in the first contrast (i.e. the first residual component, suggesting essential unidimensionality. However, the Rasch model analysis explained 29.1% of the score variance, less than the factor analytical solution.

Factor analysis

Factor analysis model showed a fair fit to the data, with RMSEA = 0.039 (90% CI: 0.000 - 0.058), CFI = 0.943 (95% CI: 0.875 - 1.190), and GFI = 0.911 (95% CI: 0.882 - 0.959). Factor loadings ranged from 0.213 to 0.822, with most items showing moderate to high loadings [Appendix 3 (Appendix 3: https://www.actamedicaportuguesa.com/revista/index.php/amp/article/view/23567/15834)].

IRT analysis

The 2-PL analysis outperformed the Rasch model, as indicated by lower AIC, SABIC, HQ, and BIC values (Table 2). This suggests that the 2-PL model provides a significantly better fit, with a more favorable balance between model complexity and predictive accuracy. This conclusion is further supported by the significant chi-square value. Overall, the results from both models reinforce the validity of the scale based on its internal structure.

Difficulty and discrimination are key item parameters in IRT. According to Baker *et al*, item discrimination values of 0.01 - 0.34 are considered very low; 0.34 - 0.64: low; 0.65 - 1.34: moderate; 1.35 - 1.69: high; and 1.70 and above very high.[30] Difficulty represents the ability level at which an item has a 50% probability of being answered correctly, with higher values indicating more challenging items. Discrimination, on the other hand, reflects how well an item distinguishes between individuals of different ability levels, with higher values indicating that the item effectively differentiates between high and low performers.

In the Rasch model, item difficulty parameters range from -4.02 (item 3) to -0.587 (item 14), reflecting a wide distribution, suggesting that the Minho-SAS possesses items with progressive difficulty levels, which can be useful if we intend to assess different levels of proficiency. In the 2-PL analysis, the item difficulty values range from -4.388 (item 1) to -0.315 (item 14), corroborating Rasch's analysis. As for discrimination parameters, ranging from 0.525 (item 1) to 2.935 (item 14), items with higher values are more effective in distinguishing suture skills and contribute more to the Minho-SAS accuracy (Table 3). In this case, we have a broad spectrum of low to very high discrimination values, according to current literature.[31]

The item characteristic curves (ICCs) and information curves (IIC) for each item as estimated by the Rasch model and the 2-PL are depicted in Fig. 1. Each curve illustrates the probability of a correct response to the item at different

Table 2 – Model fit indices for Rasch and 2PL analyses of the Minho-SAS

	AIC	SABIC	HQ	BIC	logLik	χ^2	df	*p*
Rasch	3282.212	3288.997	3305.310	3339.727	-1625.106	NA	NA	NA
2PL	3261.493	3274.215	3304.802	3369.334	-1600.746	48.7191	14	< 0.0001

AIC: Akaike information criterion; SABIC: sample-size adjusted Bayesian information criterion; HQ: Hannan-Quinn information criterion; BIC: Bayesian information criterion; Df: degrees of freedom

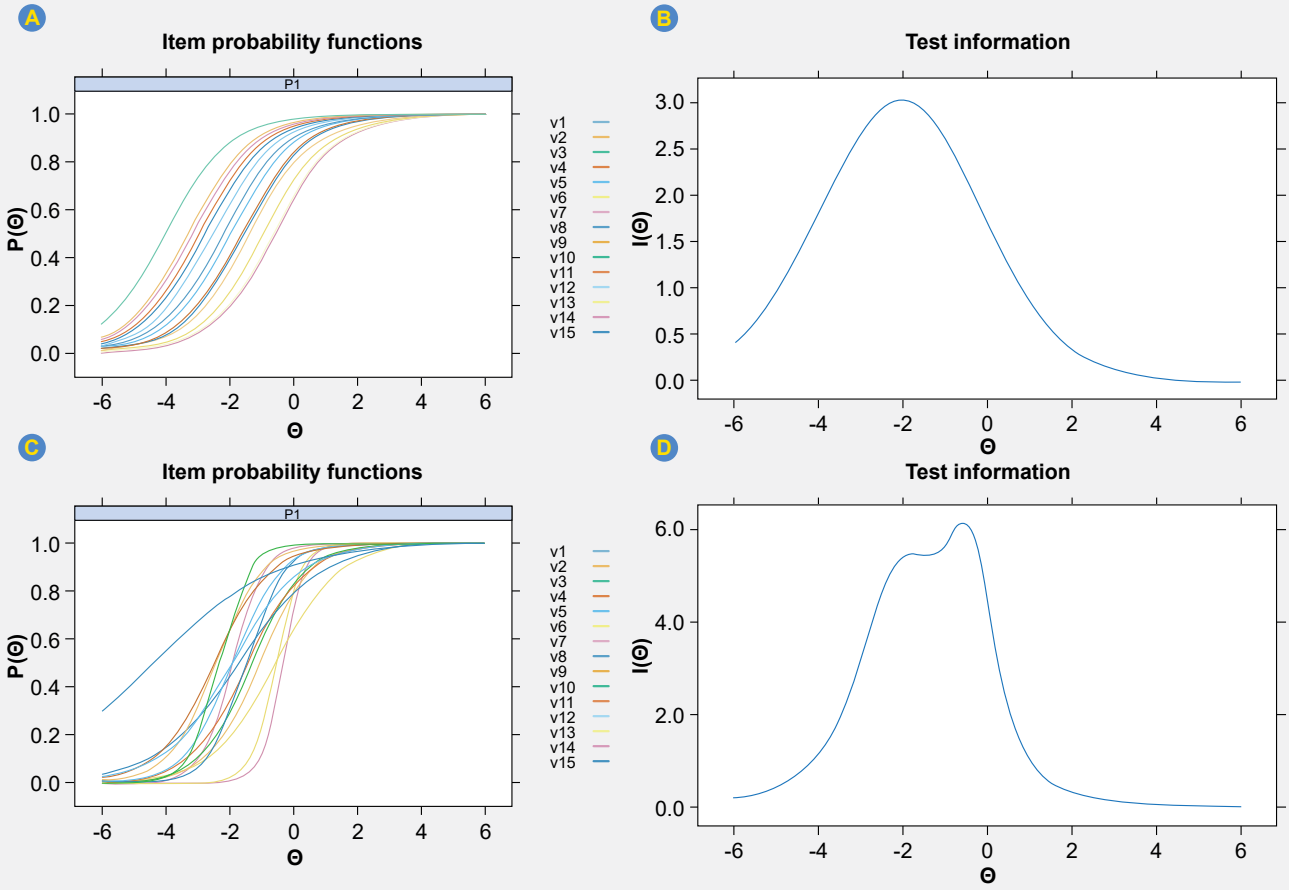

Figure 1 – Item characteristic and information curve for Rasch model (A and B) and for 2-PL model (C and D)

levels of the latent trait. The x-axis represents the latent trait continuum, indicating increasing levels of proficiency, while the y-axis shows the probability of endorsing the item. In these curves, it can be observed that, given a student whose value of Θ (theta) = 0, which would be an average student, the probability of having the item correctly would be 0.6 in item 14 (v14). When the probability is -2 Θ the value is almost 0.2 and when Θ is +4 the probability would be almost 1.

In ICC curves, item 3 exhibits a tall and narrow curve, indicating that it is very precise in measuring abilities, while item 6 has a broader ICC, suggesting that it provides information across a more comprehensive range of skills. Items with higher peaks (as item 3) are well suited for discriminating between individuals with similar abilities. The peak of the ICC represents the item's difficulty; item 2 peak is at a lower ability level, indicating that it is an easy item, while item 4 is at a higher ability level, indicating a more challenging item.

As for the IIC, the amount of information obtained from the Minho-SAS has its peak around Θ = -2 at the Rasch Analysis and Θ = 0 at the 2-PL analysis, decreasing information when going out of this range of ability. This means that in the Rasch analysis, the Minho-SAS acquires the most informative measure for students with lower skill proficiency. The 2-PL analysis captures more detail in students with average ability.

Item fit statistics (Table 4) indicate how accurately the data fit the model. Outfit could be understood as a measure of how well an item matches a pattern of responses expected by the Rasch model. If the Outfit value is too high or too low, it could indicate that the item might be too difficult or too easy when compared with the prediction. Infit focuses on the consistency of responses, being more sensitive to the pattern of responses. If the value of infit is too high or too low, it should raise awareness that the item could be confusing or not aligned with the others. Fit values should be between 0.5 to 1.5,[32] and values > 2.0 should be excluded.[33,34] In our work, we observed that the outfit statistics range from 0.479 to 1.253, which indicates acceptable variability in the behavior of the items. Similarly, the infit statistics range from 0.954 to 1.247, suggesting good levels of

Table 3 – Difficulty and discrimination parameters for each item in Rasch and 2-PL models

	Rasch model	2-PL model	
	Difficulty	Difficulty	Discrimination
1	-2.8	-4.388	0.525
2	-3.32	-2.445	1.38
3	-4.02	-2.35	2.16
4	-2.95	-2.469	1.14
5	-1.98	-1.923	0.919
6	-0.948	-0.53	2.52
7	-3.18	-1.906	2.06
8	-1.55	-1.715	0.774
9	-1.37	-1.07	1.25
10	-1.63	-1.289	1.23
11	-1.63	-1.401	1.08
12	-2.52	-1.922	1.31
13	-0.628	-0.566	1.013
14	-0.587	-0.315	2.935
15	-2.21	-1.452	1.7

Table 4 – Item fit statistics for Rasch model

	outfit	infit
Minimum	0.479	0.954
Maximum	1.253	1.247
Mean	0.782	0.971
Standard Deviation	0.195	0.118

response consistency. The mean values for outfit (0.782) and infit (0.971) indicate an overall alignment of items with the Rasch model, though with some variability in their fit.

The item-person map (Fig. 2) illustrates the distribution of participants' latent scores (top portion) and item difficulty parameters (bottom portion). Most participants are concentrated on higher levels of ability, and most items are clustered in the lower-to-middle difficulty range, with only a few items positioned at higher difficulty levels. This aligns with the assumption and goal that the Minho-SAS is a better assessment method to detect those who are not proficient.

Reliability

The scale demonstrated good internal consistency, as evidenced by McDonald's ω estimate of 0.776 (95% CI: 0.736 - 0.815) and Cronbach's α of 0.765 (95% CI: 0.723 - 0.803). These values suggest a robust level of reliability, indicating that the items comprising the scale reliably measure the intended construct. The reliability coefficient obtained for the Rasch model (0.71) and 2-PL model (0.74) reflects an acceptable range for assessment.[31]

The conditional reliability plots obtained from both the Rasch and the 2PL analyses (Fig. 3) give insights into the measurement precision of Minho-SAS across varying ability levels. In the Rasch model, the reliability is moderate level (around 0.4) for individuals with extremely low abilities (around -6 Θ). As abilities progress towards -3 to -0.5 Θ, the reliability significantly increases, stabilizing at a plateau of around 0.8. This indicates that the Rasch model is robust in measuring abilities consistently within this range. On the other hand, the 2-PL model demonstrates a notably improved trend but starts with lower reliability (less than 0.2) for those with the lowest abilities (-6 Θ). As abilities increase, the reliability swiftly improves, exceeding 0.8 around -2 Θ, and maintains this high precision up to Θ = 0. This shows that the 2-PL model consistently and accurately measures abilities at these proficiency levels. These distinct reliability trends across ability levels highlight the validation mechanisms' effectiveness in ensuring precise measurement and offer valuable guidance for future refinements in assessment methodologies.

Validity based on relations with other variables

As for the correlation between the latent traits (theta), which represents the student ability, and the sum of the items from the Minho-SAS, there was an extremely high correlation level of 0.96 from 2PL and 0.982 from the Rasch model (Table 5).

The strong positive correlation between the Rasch and

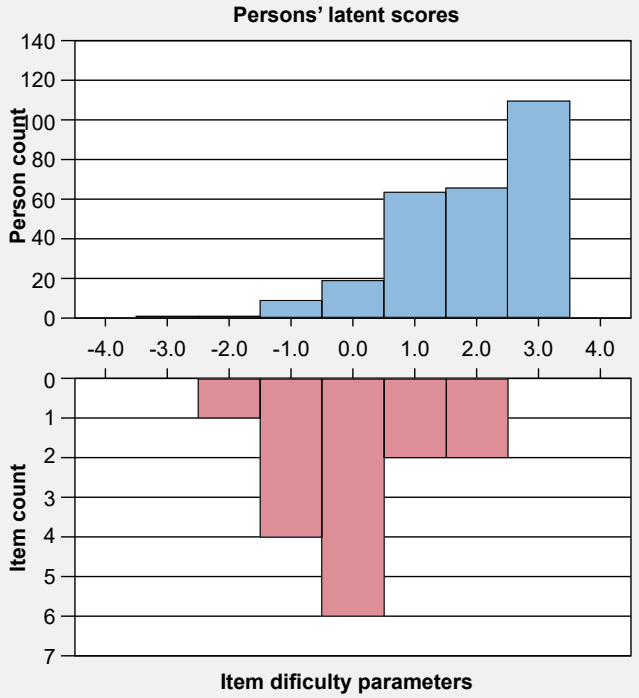

Figure 2 – Item-person map of Minho-SAS: distribution of participants' latent scores and item difficulty parameters

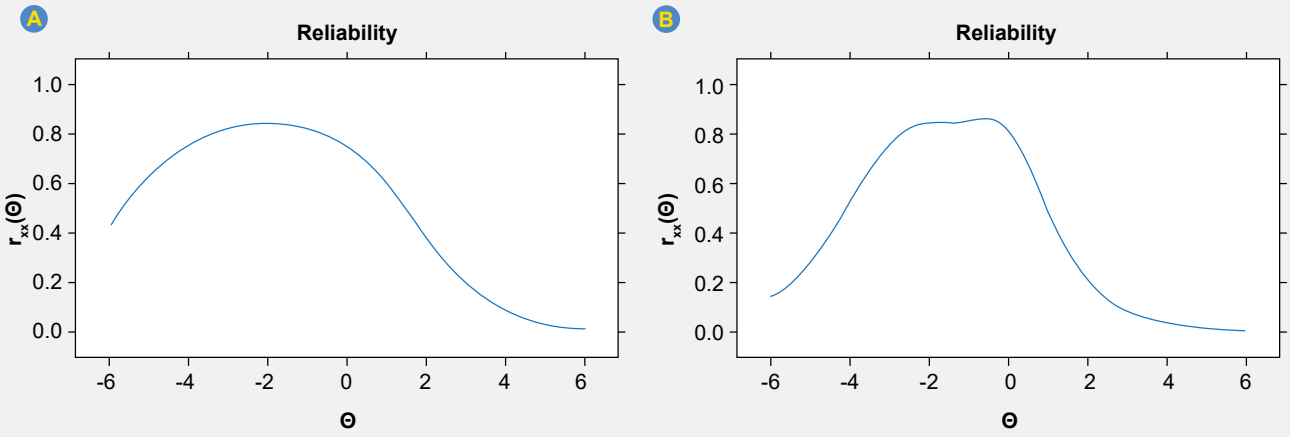

Figure 3 – Conditional reliability plot for Rasch (A) and (B) 2-PL analysis

2PL models (r = 0.975, $p < 0.001$) indicates a high agreement in measuring suturing skills. Additionally, their high correlations with the sum of Minho-SAS scores (r = 0.982 and r = 0.960, $p < 0.001$) reinforce the construct validity of the scale, as higher ability estimates align with higher total scores.

Moderate positive correlations between Minho-SAS scores and OSATS (Rasch: r = 0.670, 2PL: r = 0.698, $p < 0.001$) support its validation. However, an informatic error during the OSCE reduced the sample size (df = 135) for this analysis. The slightly lower correlations may reflect differences in item count and skill dimensions, with Minho-SAS capturing additional aspects of suturing skills. These strong associations further reinforce its validity as a robust assessment tool.

When searching for differences between the two assessed years, we conducted an independent samples *t*-test and a search for effect size using Cohen's d coefficient (Table 6).

The comparison between the assessed years reveals no difference in the Minho-SAS scores, suggesting that the scale demonstrates consistency across different cohorts of students. The observed effect sizes (Cohen's d) of 0.34 to 0.44, representing small to moderate effects, support this finding. Interestingly, there is a higher effect size on the global scale scores than the latent scores.

DISCUSSION

Suturing skills are essential for all medical graduates, making their proper assessment crucial. The Minho-SAS addresses a specific gap in surgical skill assessment by focusing exclusively on basic suturing competencies for medical students. The development process, involving multi-specialty surgical experts, ensured content validity while avoiding specialty-specific bias. The choice of a dichotomous format over a Likert scale was informed by evidence that checklist assessments can match the precision of rating scales for specific technical skills while potentially improving feedback quality and ease of use. Current literature suggests that checklist scales can be used for assessing competencies without loss of measurement precision when compared to overall rating scales, especially when referring to specific competencies, such as suturing, and can even have advantages such as improving the quality of feedback and being more user-friendly.[35,36]

The item response theory offers critical advantages for scale validation, as it models the relationship between an individual's latent trait (in this case, suturing skill) and their response to the individual items of Minho-SAS. This contrasts with the classical test theory, which considers all items as equivalents, not considering item difficulty and discrimination. For this investigation, we employed two IRT models, the Rasch model and the 2-PL, which are well-suited for validating dichotomous assessment scales. The Rasch model is a one-parameter model that assumes unidimensionality and focuses on a single latent trait, not influenced by other factors. Each item in this model has a difficulty parameter (b) that represents the point on the ability

Table 6 – Effect sizes (Cohen's d) for Minho-SAS

		Statistics	gl	*p*	Cohen's d
Minho-SAS global score	Studen's *t*	3.64	267	<0.001	0.443
Sum of Minho-SAS	Studen's *t*	2.80	267	0.005	0.342
2PL	Studen's *t*	2.93	267	0.004	0.357
Rasch	Studen's *t*	2.98	267	0.003	0.363

continuum at which individuals have a 50% probability of having a positive assessment. The 2-PL model is an extension of one-parameter models, having a discrimination (a) parameter, which allows quantifying the item's ability to discriminate between students with different ability levels.

The unidimensional structure confirmed by both factor analysis and Rasch model analysis suggests the scale measures a single, coherent construct of basic suturing ability. The wide range of item difficulties and discrimination parameters indicates the scale can differentiate between students across different proficiency levels, though it performs optimally at distinguishing lower to moderate skill levels. This aligns with the scale's primary purpose of identifying students who have not yet achieved basic competence.

The correlation between Minho-SAS and OSATS scores supports concurrent validity while highlighting their complementary nature. The OSATS provides a broader assessment of surgical skills, whereas Minho-SAS offers a detailed evaluation specific to basic suturing. The moderate correlation coefficients suggest these tools measure related but distinct aspects of surgical competence. The absence of time-based metrics in Minho-SAS represents a deliberate choice to focus on technique quality rather than speed, which may be more appropriate for assessing foundational competences.[21]

This study has several limitations. First, the sample was drawn from a single institution, which may limit the generalizability of the findings to the wider medical student population. Second, the high success rates observed across items suggest that the scale may have limited sensitivity in differentiating among highly proficient students. Nonetheless, this potential ceiling effect is less concerning given that the primary aim of the Minho-SAS is to identify minimum competence rather than to discriminate levels of excellence. Future versions of the scale might consider incorporating more challenging items to assess advanced performance, though such changes would need to remain aligned with the instrument's original purpose. Finally, the use of a single rater per student represents a methodological limitation. While rater training was implemented to mitigate variability, future applications should incorporate dual or multiple independent raters to enhance reliability and provide more robust evidence of inter-rater agreement.

The statistical evidence supports the validity and reliability of Minho-SAS for its intended purpose. The combination of different analyses provides complementary evidence for the scale's psychometric properties. The strong correlations between different scoring methods (raw scores, Rasch measures, and 2-PL estimates) indicate scoring consistency, while the conditional reliability analyses identify the ability ranges where the scale provides most precise measurement.

CONCLUSION

This study offers robust validity evidence supporting the Minho-SAS as a reliable instrument for assessing basic suturing skills in medical students. Psychometric analyses support a unidimensional internal structure and demonstrate strong reliability across various models, particularly in effectively identifying students who have yet to attain basic competences. While the scale is less precise at higher proficiency levels, this aligns with its primary purpose of certifying basic competences rather than advanced skills. Further studies are encouraged to explore its performance across diverse educational contexts and to investigate its potential use for formative assessment and targeted feedback. The Minho-SAS represents a meaningful step toward evidence-based assessment of core surgical skills, offering a solid foundation for future refinement and broader application.

ACKNOWLEDGMENTS

The authors have declared that no AI tools were used during the preparation of this work.

AUTHOR CONTRIBUTIONS

NSG, RMS: Study design, writing and critical review of the manuscript.

CC, VHP, JMP, MBA: Study design, critical review of the manuscript.

All authors approved the final version to be published.

PROTECTION OF HUMANS AND ANIMALS

The authors declare that the procedures were followed according to the regulations established by the Clinical Research and Ethics Committee and to the Helsinki Declaration of the World Medical Association updated in October 2024.

DATA CONFIDENTIALITY

The authors declare having followed the protocols in use at their working center regarding patients' data publication.

CONFLICTS OF INTEREST

The authors have no conflicts of interest to declare.

FUNDING SOURCES

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

1. Ten Cate O, Khursigara-Slattery N, Cruess RL, Hamstra SJ, Steinert Y, Sternszus R. Medical competence as a multilayered construct. Med Educ. 2024;58:93-104.
2. Epstein RM, Hundert EM. Defining and assessing professional competence. JAMA. 2002;287:226.
3. Witheridge A, Ferns G, Scott-Smith W. Revisiting Miller's pyramid in medical education: the gap between traditional assessment and diagnostic reasoning. Int J Med Educ. 2019;10:191.
4. Frank JR, Danoff D. The CanMEDS initiative: implementing an outcomes-based framework of physician competencies. Med Teach. 2007;29:642-7.
5. Moskowitz EJ, Nash DB. Accreditation council for graduate medical education competencies: practice-based learning and systems-based practice. Am J Med Qual. 2007;22:351-82.
6. Ten Cate O. Entrustability of professional activities and competency-based training. Med Educ. 2005;39:1176-7.
7. Epstein RM. Assessment in medical education. Cox M, Irby DM, editors. N Engl J Med. 2007;356:387-96.
8. Emmanuel T, Nicolaides M, Theodoulou I, Yoong W, Lymperopoulos N, Sideris M. Suturing skills for medical students: a systematic review. In Vivo. 2021;35:1-12.
9. Vaidya A, Aydin A, Ridgley J, Raison N, Dasgupta P, Ahmed K. Current status of technical skills assessment tools in surgery: a systematic review. J Surg Res. 2020;246:342-78.
10. Bilgic E, Endo S, Lebedeva E, Takao M, McKendy KM, Watanabe Y, et al. A scoping review of assessment tools for laparoscopic suturing. Surg Endosc. 2018;32:3009-23.
11. Almeland SK, Lindford A, Sundhagen HP, Hufthammer KO, Strandenes E, Svendsen HL, et al. The effect of microsurgical training on novice medical students' basic surgical skills - a randomized controlled trial. Eur J Plast Surg. 2020;43:459-66.
12. Dormegny L, Neumann N, Lejay A, Sauer A, Gaucher D, Proust F, et al. Multiple metrics assessment method for a reliable evaluation of corneal suturing skills. Sci Rep. 2023;13:2920.
13. Nugent E, Joyce C, Perez-Abadia G, Frank J, Sauerbier M, Neary P, et al. Factors influencing microsurgical skill acquisition during a dedicated training course. Microsurgery. 2012;32:649-56.
14. Chipman JG, Schmitz CC. Using objective structured assessment of technical skills to evaluate a basic skills simulation curriculum for first-year surgical residents. J Am Coll Surg. 2009;209:364.
15. Goova MT, Hollett LA, Tesfay ST, Gala RB, Puzziferri N, Kehdy FJ, et al. Implementation, construct validity, and benefit of a proficiency-based knot-tying and suturing curriculum. J Surg Educ. 2008;65:309-15.
16. Sato E, Mitani S, Nishio N, Kitani T, Sanada T, Ugumori T, et al. Development of proficiency-based knot-tying and suturing curriculum for otolaryngology residents: a pilot study. Auris Nasus Larynx. 2020;47:291-8.
17. Thinggaard E, Zetner DB, Fabrin A, Christensen JB, Konge L. A study of surgical residents' self-assessment of open surgery skills using gap analysis. Simul Healthc. 2023;18:305-11.
18. Buckley CE, Kavanagh DO, Gallagher TK, Conroy RM, Traynor OJ, Neary PC. Does aptitude influence the rate at which proficiency is achieved for laparoscopic appendectomy? J Am Coll Surg. 2013;217:1020-7.
19. Nickel F, Brzoska JA, Gondan M, Rangnick RM, Chu J, Kenngott HG, et al. Virtual reality training versus blended learning of laparoscopic cholecystectomy: a randomized controlled trial with laparoscopic novices. Medicine. 2015;94:e764.
20. Ebina K, Abe T, Hotta K, Higuchi M, Furumido J, Iwahara N, et al. Objective evaluation of laparoscopic surgical skills in wet lab training based on motion analysis and machine learning. Langenbecks Arch Surg. 2022;407:2123-32.
21. Sundhagen HP, Almeland SK, Hansson E. Development and validation of a new assessment tool for suturing skills in medical students. Eur J Plast Surg. 2018;41:207-16.
22. Vassiliou MC, Feldman LS, Andrew CG, Bergman S, Leffondré K, Stanbridge D, et al. A global assessment tool for evaluation of intraoperative laparoscopic skills. Am J Surg. 2005;190:107-13.
23. De Champlain AF. A primer on classical test theory and item response theory for assessments in medical education. Med Educ. 2010;44:109-17.
24. Bond T, Yan Z, Heene M. Applying the Rasch model: fundamental measurement in the human sciences. 4th ed. New York: Routledge; 2020.
25. Tavakol M, Dennick R. The foundations of measurement and assessment in medical education. Med Teach. 2017;39:1010-5.
26. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for Educational and Psychological Testing. 2014. [cited 2024 Feb 06]. Available from: https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf.
27. Boateng GO, Neilands TB, Frongillo EA, Melgar-Quiñonez HR, Young SL. Best practices for developing and validating scales for health, social, and behavioral research: a primer. Front Public Health. 2018;6:1-18.
28. Goodhue DL, Lewis W, Thompson R. Does PLS have advantages for small sample size or non-normal data? MIS Q. 2012;36:981-1001.
29. Martin JA, Regehr G, Reznick R, MacRae H, Murnaghan J, Hutchison C, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. Br J Surg. 1997;84:273-8.
30. Baker FB. The basics of item response theory. 2nd ed. 2001. [cited 2025 Mar 23]. Available from: https://eric.ed.gov/?id=ED458219.
31. Walsh A, Cao R, Wong D, Kantschuster R, Matini L, Wilson J, et al. Using item response theory (IRT) to improve the efficiency of the Simple Clinical Colitis Activity index (SCCAI) for patients with ulcerative colitis. BMC Gastroenterol. 2021;21:132.
32. Institute for Objective Measurement. What do infit and outfit, mean-square and standardized mean? [cited 2023 Nov 16]. Available from: https://www.rasch.org/rmt/rmt162f.htm.
33. Cecilio-Fernandes D, Medema H, Collares CF, Schuwirth L, Cohen-Schotanus J, Tio RA. Comparison of formula and number-right scoring in undergraduate medical training: a Rasch model analysis. BMC Med Educ. 2017;17:192.
34. Institute for Objective Measurement. Dichotomous mean-square fit statistics. [cited 2023 Nov 16]. Available from: https://www.rasch.org/rmt/rmt82a.htm.
35. Wood TJ, Pugh D. Are rating scales really better than checklists for measuring increasing levels of expertise? Med Teach. 2020;42:46-51.
36. Ilgen JS, Ma IW, Hatala R, Cook DA. A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. Med Educ. 2015;49:161-73.