Original Article

Journal of Advances in Medical Education & Professionalism

Surgical Professionalism: Development and Educational Suitability Assessment of a Combined Situational Judgment Test and Guided Reflection among General Surgery Residents in Indonesia

DANIEL ARDIAN SOESELO[1,2*], PhD Candidate; RENNIE YOLANDA[3], MD; PATRICIA WINONA[3], MD; YUNISA ASTIARANI[4], MPH; TOAR JEAN MAURICE LALISANG[5], PhD; RETNO ASTI WERDHANI[6], PhD; AGUS PURWADIANTO[7], PhD; WIRSMA ARIF HARAHAP[8], PhD; THEDDEUS OCTAVIANUS HARI PRASETYONO[9, 10], PhD; DIANTHA SOEMANTRI[11], PhD; ARDI FINDYARTINI[11], PhD

[1]Department of Surgery and Medical Education Unit, School of Medicine and Health Sciences, Atma Jaya Catholic University of Indonesia, Jakarta, Indonesia; [2]Doctoral Programme in Medical Sciences, Faculty of Medicine, Universitas Indonesia, Jakarta, Indonesia; [3]Department of Surgery, School of Medicine and Health Sciences, Atma Jaya Catholic University of Indonesia, Jakarta, Indonesia; [4]Department of Public Health and Nutrition, School of Medicine and Health Sciences, Atma Jaya Catholic University of Indonesia, Jakarta, Indonesia; [5]Department of Surgery, Cipto Mangunkusumo Hospital, Faculty of Medicine, Universitas Indonesia, Jakarta, Indonesia; [6]Department of Community Medicine, Faculty of Medicine, Universitas Indonesia, Jakarta, Indonesia; [7]Department of Forensic Medicine and Medicolegal Studies, Cipto Mangunkusumo Hospital, Faculty of Medicine, Universitas Indonesia, Jakarta, Indonesia; [8]Department of Surgery, Faculty of Medicine, Universitas Andalas, Padang, Indonesia; [9]Division of Plastic Surgery, Department of Surgery, Cipto Mangunkusumo Hospital, Faculty of Medicine, Universitas Indonesia, Jakarta, Indonesia; [10]Indonesian Clinical Training and Education Center (ICTEC), Cipto Mangunkusumo Hospital, Jakarta, Indonesia; [11]Department of Medical Education and Medical Education Center, Indonesian Medical Education and Research Institute (IMERI), Faculty of Medicine, Universitas Indonesia, Jakarta, Indonesia

Abstract

Introduction: Assessing professionalism in a valid, objective, and cost-effective manner remains a persistent challenge in clinical education. In Indonesia, this issue is particularly relevant in surgical training, where residents must integrate technical competence with professional behaviour in high-stakes, resource-limited settings. This study aimed to develop and evaluate the educational suitability of a Situational Judgement Test (SJT) combined with guided written reflection to assess and promote professionalism among general surgery residents.

Methods: A mixed-methods sequential design was used. An SJT was developed through blueprinting, item writing, and expert validation, resulting in 26 scenarios across six professionalism domains. Residents (n=64 at baseline; n=44 completers) completed a proctored pre-test SJT, participated in an 8-week guided reflection program (four submissions), and completed an asynchronous post-test SJT. Educational suitability evidence was examined through content validity (CVI) and internal consistency (Kendall's W). No construct or criterion validity testing was performed. Reflections were scored using Kember's four-level reflective framework by three independent assessors with Fleiss' Kappa for inter-rater agreement. Changes in SJT and reflection depth were analysed using paired tests and correlation analysis. Data were analysed using SPSS version 23.

Results: The mean CVI across items was 0.97 (range 0.90–1.00). The mean SJT score increased significantly from 386.59±19.31 to 420.45±12.56 among completers, with a mean difference of 33.86±16.03 ($p<0.001$). Reliability improved from Kendall's W=0.38 to 0.52. Sensitivity analysis comparing baseline pre-test scores between completers (386.6±19.3) and dropouts (391.0±15.5) showed no significant difference (Mann–Whitney U, $p=0.258$), indicating limited attrition bias. Reflective depth progressed from median level 1 to 4 across four submissions (median $\Delta=+3$, $p<0.001$), with substantial inter-rater agreement ($\kappa=0.68$). There was no significant correlation between Δ reflection and Δ SJT ($r=0.059$, $p=0.703$).

Conclusions: Integrating SJT with guided reflection demonstrated strong content validity, internal consistency, and educational benefit, supporting its suitability for formative educational use in surgical professionalism training.

Keywords: Professionalism, Reflection, Medical professionalism, Surgeon

*Corresponding author:
Daniel Ardian Soeselo, PhD Candidate;
Postal address: Jl. Pluit Raya No.2 21, RT.21/RW.8, Penjaringan, Kec. Penjaringan, Jkt Utara, Postal code: 14440, Daerah Khusus Ibukota, Jakarta, Indonesia
Tel: +62-81381937250
Email: daniel.ardian@atmajaya.ac.id

Please cite this paper as: Soeselo DA, Yolanda R, Winona P, Astiarani Y, Lalisang TJM, Werdhani RA, Purwadianto A, Harahap WA, Prasetyono TOH, Soemantri D, Findyartini A. Surgical Professionalism: Development and Educational Suitability Assessment of a Combined Situational Judgment Test and Guided Reflection among General Surgery Residents in Indonesia. J Adv Med Educ Prof. 2026;14(3):263-271. DOI: 10.30476/jamp.2026.109158.2298.

Received: 11 October 2025
Accepted: 19 April 2026

Introduction

Professionalism is a core competency emphasized in both the Accreditation Council for Graduate Medical Education (ACGME) framework and the Indonesian General Surgery Curriculum. This competency encompasses values, attitudes, and behaviours that reflect integrity, accountability, empathy, and a commitment to patient care and continuous improvement (1–4). In the dynamic and high-pressure environment of surgical training, the cultivation and reinforcement of professionalism are increasingly critical to ensure the quality of care and patient safety (5–7).

However, assessing and enhancing professionalism remains a significant challenge in medical education, particularly in Indonesia. Traditional assessment tools often fail to capture the complexities of real-world clinical situations encountered by surgical residents (8, 9). In the Indonesian residency setting, these challenges are compounded by contextual factors such as heavy clinical workloads due to uneven workforce distribution, wide variability in supervision quality across teaching hospitals, and the limited availability of validated instruments to formally assess professionalism (9, 10). These conditions create gaps in both monitoring and fostering professional behaviour during training.

To address these challenges, there is a growing need for more contextualized and reflective approaches to assess and promote professionalism. The Situational Judgment Test (SJT) has been widely implemented internationally as a method to evaluate non-cognitive attributes such as professionalism, empathy, and ethical decision-making within realistic clinical scenarios (11–13).

In addition, guided written reflection has been recognized as an effective educational strategy to help residents internalize professional values through personal experience (14-16). Through guided reflection, residents are encouraged to critically examine their thoughts, feelings, and actions in response to complex clinical encounters, facilitating deeper self-awareness and growth (16, 17).

Integrating the SJT as an assessment tool with written reflection as a pedagogical approach offers a potentially synergistic model—allowing not only the measurement but also the reinforcement of professionalism through both cognitive and experiential learning pathways.

This study, therefore, aims to develop and evaluate an educational intervention combining SJT and guided written reflection to enhance medical professionalism among general surgery residents in Indonesia. It is anticipated that this approach will provide deeper insights into the effectiveness of these methods and contribute to the development of a more contextualized and impactful professionalism curriculum for surgical residency programs.

Methods

Study Design

A mixed-methods design was employed to develop, pilot, and validate a combined SJT and guided written reflection as tools for assessing and enhancing professionalism among general surgery residents in Indonesia. The quantitative component involved SJT development, administration, and pre–post analysis of scores, while the qualitative component involved evaluation of guided reflections on professional experiences.

Situational Judgement Test (SJT) Development

The development process of the Situational Judgment Test (SJT) began with the creation of a test blueprint, which outlined the purpose, format, target population, and professionalism attributes to be assessed. The primary purpose of the assessment was to provide a formative evaluation of professionalism among general surgery residents. The assessment format was designed as a written, scenario-based test targeted at residents from their second year through the final year of training. Six professionalism domains were identified, drawing from the Indonesian General Surgery Curriculum (3), the Accreditation Council for Graduate Medical Education (ACGME) framework (2), the Indonesian Medical Code of Ethics (18), and findings from a national study on surgical professionalism in Indonesia (4). These domains included accountability in completing duties as a physician, integrity, clinical decision-making, effective communication, respect for collegial relationships, and teamwork. Each scenario presented a professional dilemma, and participants were required to rank five possible responses from 1 “most appropriate” to 5 “least appropriate”, assigning each rank only once.

To ensure item quality and alignment, we recruited seven general surgeons with at least three years of clinical experience as item writers. They participated in a one-day training workshop (July 2024) on SJT design, including principles of scenario development, response format, and linkage to professionalism attributes. Over the subsequent two months, the team produced a pool of 62 scenarios, which were then reviewed for clarity, context, and relevance. Scenarios with fewer than five response options or unclear dilemmas were revised or discarded, resulting in 49 refined scenarios for expert review.

Table 1. Distribution of Selected SJT Scenarios by Professionalism Attribute

No.	Professionalism Attribute	Number of Scenarios (N=26)
1	Accountability in completing duties as a physician	4
2	Integrity	3
3	Decision making	6
4	Effective communication	5
5	Respect for collegial relationships	6
6	Teamwork	2

Content Validation and Answer Key Development

A panel of 10 subject matter experts (SMEs)—comprising seven senior surgeons and three experts in medical ethics or law, each with at least five years of experience—conducted content validation of the scenarios. SMEs evaluated each scenario for relevance to the intended professionalism domain and collaboratively established consensus-based answer keys. Content validity was quantified using the Content Validity Index (CVI) at both the item and scale levels (19). Scenarios with inconsistent ratings, overlapping content, or lack of consensus were excluded. Based on this process, 23 scenarios were eliminated, leaving 26 validated scenarios spanning six professionalism domains (Table 1).

The SJT answer key was developed through a modified Delphi process involving 10 expert surgeons and medical educators. Experts independently ranked the response options for each item and subsequently participated in two consensus rounds. Agreement thresholds of ≥70% were used to finalize rank orders. This procedure ensured standardized and transparent scoring across test items.

SJT Scoring System

Each scenario contained five possible actions ranked by participants from most to least appropriate. Scoring was based on the degree of deviation from the expert consensus key. Exact match scored 4 points, one-rank deviation scored 3 points, two-rank deviation scored 2 points, three-rank deviation scored 1 point, and four-rank deviation scored 0. Total scores ranged from 208 to 520 across all scenarios. The scores were summed across items, possible scores ranged from 8 to 20 per scenario, with total possible scores ranging from 208 to 520. This approach follows previous international SJT validation studies assessing gradations in professional judgment (19, 20).

Guided Written Reflection

In conjunction with the SJT, guided written reflection was introduced as a complementary formative exercise. The framework was adapted from Gibbs' Reflective Cycle, which guides the learners to describe experiences, explore emotions, evaluate outcomes, and identify lessons for future practice. Substitute by each resident submitted four reflective essays. Over an eight-week period, scheduled biweekly. Reflection prompts focused on real clinical experiences involving professional dilemmas, such as ethical conflicts, teamwork challenges, or communication issues. Reflections were independently evaluated by trained assessors (DAS, RY, PAW), using Kember's four-level reflective depth framework, classifying responses as descriptive writing, descriptive reflection, dialogic reflection, or critical reflection. This allowed progressive tracking of reflective depth across the four submissions.

Study Procedure

Participants first completed the baseline SJT in a synchronous, proctored session. 64 participants completed the assessment within a fixed 90-minute time window. They were then introduced to the reflective writing component with instructions on the structure and expectations. Over the following eight weeks, they submitted their four reflections electronically, which were anonymized for assessment. After completing all reflections, 44 participants took the same SJT again as an asynchronous post-test independently within a 2-week period to measure the changes in professional judgment.

Participants and Sampling

Participants were general surgery residents from the University of Indonesia and Diponegoro University. Eligibility criteria included being in at least the third semester of residency and having at least six months of training remaining at the time of recruitment. Residents who did not complete the post-test or failed to submit all required written reflections were excluded.

Sample size estimation was based on the formula for paired numerical analytic studies:

$$n = \left(\frac{(Z_\alpha + Z_\beta) \cdot S}{X_1 - X_2}\right)^2$$

where:

Z_{α}=1.96 for α=0.05
Z_{β}=1.28 for β=0.10 (power 90%)
S=estimated SD of paired differences, set as twice the minimal meaningful difference
X_1-X_2=15 points (minimal detectable mean difference)

After adjusting for a 20% anticipated attrition rate, the required sample size was 44 participants.

Data Analysis

Quantitative data (SJT scores) were analysed using paired statistical tests (paired T-test or Wilcoxon signed-rank test, depending on normality) to compare pre- and post-test performance. Reliability was examined using Kendall's W coefficient of concordance. Sensitivity analysis compared baseline pre-test scores between completers and dropouts (Mann–Whitney U used when distribution non-normal).

Qualitative data from the guided written reflections were analysed using a structured approach based on Kember's four-level reflective depth framework (21). To ensure methodological transparency, the analysis process followed four steps:

1. Framework Operationalization: Kember's levels (descriptive writing, descriptive reflection, dialogic reflection, and critical reflection) were translated into an analytic rubric with explicit indicators for content, reasoning, emotional insight, and demonstrated learning. Each reflection was assigned a score from 1 to 4 corresponding to these levels.

2. Independent Rating by Three Assessors: Three trained assessors (DAS, RY, PAW) independently reviewed each anonymized reflection. Each assessor coded the narrative according to the rubric without awareness of the participant's identity or time of submission.

3. Consensus and Discrepancy Resolution: Coding sheets were compared, and disagreements were discussed in structured adjudication meetings. Consensus scores were determined when at least two assessors agreed. When all three ratings differed, the assessors re-reviewed the reflection and reached final consensus through discussion, ensuring interpretive rigor.

4. Reliability and Progression Analysis: Inter-rater reliability was assessed using Fleiss' Kappa. Median reflection scores were calculated for each of the four submission rounds to evaluate developmental progression over time. Changes in reflection depth (Δ reflection) were then compared with changes in SJT performance (Δ SJT), using correlation analysis.

All qualitative analyses were conducted manually using structured coding matrices. This process ensured analytic transparency, reproducibility, and alignment with established reflective writing research methodologies. In addition, an exploratory cluster analysis (k – means) was conducted to identify the patterns of participant development based on two variables: change in reflective depth (Reflection 4 – Reflection 1) and change in SJT score (post-test – pre-test). Clustering was performed at the case level (individual participants) to illustrate heterogeneity in professionalism learning.

Statistical analyses were performed using SPSS version 23.

Ethical Consideration

We obtained ethical approval from The Ethics Committee of the Faculty of Medicine, University of Indonesia, Cipto Mangunkusumo Hospital number KET-741/UN2.F1/ETIK/PPM.00.02/2024.

Results

Demographic Characteristics of the Respondents

A total of 64 residents were recruited, of whom 44 (68.7%) completed the study protocol. Twenty participants (31.3%) were excluded due to dropout, primarily because of incomplete reflection submissions or absence at post-test. Among the 44 completers, 24 were from University of Indonesia and 20 from Diponegoro University. Most of the participants were male (68.2%). 24 were in the intermediate phase of training (semesters 4–7), while the remainder were in advanced phases (semester ≥8).

Content Validation of SJT Scenarios

All 26 SJT scenarios were reviewed by a panel of subject matter experts (SMEs) to assess content validity. The average Content Validity Index (CVI) across all items was 0.97, ranging from 0.90 to 1.00. Twenty-four scenarios (92.3%) achieved a perfect CVI score of 1.00, while two scored 0.90. Based on SME feedback, minor wording adjustments were made to enhance clarity, but no scenario was excluded. These results indicate that the SJT items demonstrated strong content validity and were suitable for pilot testing.

Reliability and Changes in SJT Scores

Reliability of the SJT, measured using Kendall's W (20), improved from moderate (W=0.38) at pre-test to substantial (W=0.52) at post-test. The mean total SJT score increased significantly from 386.6±19.3 to 420.5±12.6, with a mean difference of 33.9±16.0 ($p<0.001$). The magnitude of improvement yielded a large, standardized effect size (Cohen's d_z=2.11, 95% CI: 1.81–2.42).

Table 2. Pre- and Post-Test SJT Scores by Professionalism Domain (N=44)

Category	Pre-test (Mean±SD)	Post-test (Mean±SD)	Δ (Mean±SD)	p-value
Total SJT score	386.59±19.31	420.45±12.56	33.86±16.03	<0.001*
Domain scores				
Accountability	60.68±6.59	67.73±4.06	7.04±5.98	<0.001*
Integrity	51.32±3.68	53.64±3.02	2.31±3.55	<0.001*
Clinical decision-making	84.23±8.28	94.09±6.08	9.86±7.56	<0.001**
Effective communication	74.86±6.38	79.36±6.53	4.50±5.67	<0.001**
Collegial respect	84.64±5.57	91.86±6.88	7.22±6.90	<0.001*
Teamwork	30.86±4.23	32.41±3.18	1.54±4.02	0.014*

*Wilcoxon signed-rank test; **Paired T-test

Table 3. Reflective Depth Scores (Kember Framework)

Reflection	Median score (Min–Max)	Δ Median	p-value
Reflection 1	1 (1–4)	+ 3	<0.001
Reflection 2	2 (1–4)		
Reflection 3	3 (1–4)		
Reflection 4	4 (1–4)		

Table 4. Cluster Analysis of Reflection and SJT Outcomes

Cluster group	n	Reflection change	Post-test SJT (Mean±SD, range)	Mean Δ SJT*
Group 1	3	Remained non-reflective	438.0±17.4 (418–450)	35.3
Group 2	34	Became reflective	418.3±11.0 (386–442)	20.6
Group 3	7	Consistently reflective	423.3±13.2 (404–440)	26.4

*Mean change in SJT score

However, interpretation of this magnitude should consider the single-group pre–post design and potential influences such as practice effects and reduced post-test score variability. Sensitivity analysis showed no significant difference in the baseline SJT scores between participants who completed the study (386.6±19.3) and those who dropped out (391.0±15.5), Mann–Whitney results p=0.258. Domain-level analysis revealed significant improvements across all six professionalism attributes. The largest gain was observed in clinical decision-making (+9.9 points), followed by collegial relationships (+7.2) and accountability (+7.0). The smallest improvement was in teamwork (+1.5). Full results are presented in Table 2.

Development of Reflective Depth

The participants submitted four written reflections throughout the study period. Inter-rater agreement across the three assessors was moderate (64.2%), with Fleiss' Kappa value (κ=0.68) indicating substantial reliability. Reflection scores showed progressive improvement over time: the median score increased from 1 (descriptive writing) in the first reflection to 4 (critical reflection) in the final submission. The overall median improvement of three points (p<0.001) indicates that repeated guided reflection effectively promoted deeper critical engagement with professionalism dilemmas. Results are summarized in Table 3.

Relationship between Reflection and SJT Performance

Correlation analysis revealed no significant association between the changes in reflective depth and changes in SJT scores (r=0.059, p=0.703). While the participants consistently improved in reflective depth, this was not directly accompanied by equivalent increases in SJT performance.

Cluster Analysis of Reflection–SJT Development Patterns

To further explore individual variability in developmental trajectories, we performed a cluster analysis using changes in SJT scores (post–pre) and changes in reflective depth (Reflection 4 – Reflection 1). Three distinct participant clusters emerged:

1. Cluster 1 – Non-reflective but high SJT gain (n=3): Participants demonstrated minimal improvement in reflective depth but showed the largest increases in SJT scores.

2. Cluster 2 – Reflective with modest SJT gain (n=34): This majority group showed substantial improvement in reflective depth but only

moderate increases in SJT performance.

3. Cluster 3 – Consistently reflective with moderate SJT gain (n=7): Participants showed stable high reflection scores throughout the study while achieving moderate SJT gains.

These patterns illustrate heterogeneity in how residents integrate reflective capacity and situational judgment across the study period. The distribution of clusters is presented in Table 4.

Discussion

This study provides preliminary validity evidence supporting the educational suitability of the developed ranking-format SJT within surgical professionalism training. The high Content Validity Index and structured Delphi-based answer key development support evidence based on the content and response process. Moderate Kendall's W coefficients are consistent with the multidimensional and scenario-based characteristics of situational judgment assessments, particularly in postgraduate contexts. However, this study does not provide full structural, criterion-related, or longitudinal reliability evidence. Given the sample size, advanced psychometric analyses such as factor analysis or test–retest reliability were not feasible. Therefore, the findings should be interpreted as early-phase educational suitability evidence rather than comprehensive psychometric validation.

Relationship between SJT and Reflection

This study utilized 26 knowledge-based scenarios in a ranking response format. While other studies have adopted rating or single-best-answer formats (22, 23). The use of a ranking-format SJT allowed the participants to evaluate multiple professional options simultaneously, engaging higher-order reasoning consistent with postgraduate training. This format required the participants to rank them from most to least appropriate, aligning with the principles of knowledge-based assessment (24). The moderate baseline concordance in the participants' responses suggests diverse interpretations of professionalism dilemmas, which is expected given the multidimensional nature of SJTs and variability in clinical experience. This finding is consistent with those of Sorell, et al. (19), who reported that internal consistency for SJT assessments typically falls within the low-to-moderate range due to their multidimensional nature. The ranking response format and the diverse demographic and institutional backgrounds of the participants across Indonesia likely contributed to this variation. The post-intervention increase in concordance reflects a more aligned understanding of appropriate professional behaviour, indicating that the combined SJT–reflection intervention helped shape shared professional norms.

The lack of significant difference in baseline SJT performance between the completers and non-completers suggests that participant attrition did not occur selectively among lower-performing individuals. Therefore, the observed improvement in post-test SJT scores is unlikely to be explained by attrition bias, and the internal validity of the pre–post findings is preserved. Attrition bias is a common concern in longitudinal educational studies, particularly when follow-up assessments are asynchronous.

Although the standardized effect size was numerically large, caution is warranted in its interpretation. Educational interventions rarely produce effect sizes exceeding 1.0 under controlled conditions. The absence of a control group, possible practice effects, and the asynchronous post-test administration may have contributed to inflation of the observed magnitude. Thus, the score improvement should be interpreted as evidence of educational responsiveness rather than definitive behavioural transformation.

Reflection depth improved significantly over the study period; however, its weak association with SJT gains highlights the point that cognitive awareness of professionalism does not always translate directly into applied judgment. This distinction aligns with current theoretical perspectives that view professionalism as comprising parallel but interacting domains: cognitive, affective, and behavioural, each of which can develop at different rates.

The cluster analysis further demonstrates that residents follow different developmental trajectories. Some showed strong judgment improvement without reflective growth, suggesting reliance on experiential intuition. Others exhibited deep reflection without major SJT gains, likely representing early internalization without behavioural manifestation. These patterns emphasize that professionalism development is individualized and multifaceted. Reflection captures internal awareness of values and introspective thinking processes, whereas the SJT assesses the behavioural choices participants would make in specific situations (15, 25–27).

Cluster Analysis: Three Patterns of Participant Development

Cluster analysis was conducted to explore the diversity of the participants' responses to professionalism learning. Based on changes in reflection scores and SJT scores, three groups of

participants were identified, each demonstrating different developmental characteristics. These findings reinforce the notion that professionalism is not a single construct, but rather a set of cognitive, affective, and behavioural aspects that may evolve differently across individuals. Some participants appeared capable of producing deep written reflections, yet they were not able to translate these into clinical decision-making.

Overall, the results indicate that improvements in the quality of written reflections are not always accompanied by higher SJT scores. This gives rise to several interpretations that can be explained from the perspective of medical education theory and professionalism learning.

Educational Implications and Theoretical Interpretation

1. Reflection as a Gradual Internal Process

Reflection is a metacognitive process through which individuals analyse their experiences to construct new meaning and understanding (27). In this study, the participants in Group 2 who demonstrated higher reflection scores but limited SJT improvement may still be in the process of internalizing professional values. While their cognitive awareness of professionalism increases, their ability to enact these values in real-world decision-making may develop more slowly. Two factors that influence the reflective writing are: 1) motivation to reflect, and 2) metacognitive skills for reflection. In addition, writing skills are needed to effectively produce a reflective narrative (15, 16).

2. Differences in Cognitive Domains between Reflection and SJT

Written reflection and the SJT essentially assess different cognitive domains. Written reflection evaluates self-awareness, personal values, and metacognitive thinking, whereas the SJT assesses decision-making abilities in complex situations involving professional interactions, ethics, and communication. Therefore, discrepancies between reflection and SJT gains are expected, as each targets a distinct level of cognitive and behavioural competence.

Participants in Group 1, who demonstrated the highest SJT scores while remaining non-reflective in writing, may possess strong professional intuition or practical experience but have not yet been able to articulate these experiences in reflective narratives. Conversely, participants in Group 2 who showed reflective development may not have fully integrated their value awareness with real-time clinical behaviour.

3. Integration of Reflection and Experiential Learning

The findings underscore the need for integrated professionalism education that combines reflective writing with experiential learning. Reflection exercises should be paired with simulations, case-based discussions, and structured feedback to bridge cognitive insight and behavioural performance. Such integration promotes both self-awareness and decision-making skills essential to professionalism.

Furthermore, these findings highlight the significant influence of the hidden curriculum—the implicit social, cultural, and hierarchical norms that shape the learners' professional identity and behaviour beyond formal instruction. As described by Poola, et al. (2021), professionalism learning is often mediated through informal interactions, mentorship, and institutional culture rather than explicit teaching alone (17).

In summary, the findings support a dual-path model of professionalism development. Reflection builds the cognitive and moral foundation of professionalism, while situational judgment tests its practical enactment. Medical education should, therefore, design curricula that integrate both reflective and applied learning modalities to cultivate competent, ethical, and adaptive surgeons.

Limitations

The SJT assessment is highly influenced by the test development team and subject matter experts (SMEs). One of the main challenges for the item writers is creating dilemmas that are realistic and designing response options that are rational and likely to be chosen. The qualifications of the SMEs are also crucial in reviewing scenarios and providing the scoring key. In addition, the absence of a strict time limit for the online post-test (asynchronously) could have allowed the participants more time to deliberate, potentially inflating post-test scores.

This study employed a single-group pre–post design without a control group. Consequently, improvements in SJT scores cannot be attributed exclusively to the intervention, as practice effects, increased test familiarity, or maturation over time may have contributed to score changes. Although sensitivity analysis suggested minimal attrition bias, future studies employing randomized controlled designs or comparison groups are necessary to establish causal effects and to rule out alternative explanations.

Conclusion

The combination of SJT assessment and guided

written reflection demonstrated educational responsiveness and strong content relevance within surgical professionalism training. At this stage of development, the instrument is best positioned for formative educational use. Further research involving larger samples and expanded psychometric evaluation is required before broader implementation or high-stakes application.

Acknowledgement

We sincerely thank all participants for their valuable insights and time.

Authors' contribution

DA.S, R.Y, A.F, TO.HP and DS conceptualized and designed the study, curated and analysed the data. DA.S and A.F also supervised the process and aligned the overall manuscript. P.W and Y.A prepared the manuscript and assisted in revising the manuscript. RA.W, TJ.ML, A.P, WA.H, contributed to data collection and enriched the contextual analysis. All authors reviewed and approved the final version of the manuscript.

Conflict of interests

The authors report no conflicts of interest.

Declaration of AI Use

The authors used generative artificial intelligence (ChatGPT, OpenAI) to assist with language editing, grammar refinement, and manuscript organization. All scientific content, study design, data analysis, interpretation of findings, and final manuscript preparation were performed and verified by the authors. The authors take full responsibility for the accuracy and integrity of the work.

References

1. Meara JG, Leather AJM, Hagander L, Alkire BC, Alonso N, Ameh EA, et al. Global Surgery 2030: Evidence and solutions for achieving health, welfare, and economic development. The Lancet. 2015;386(9993):569–624.
2. Accreditation Council for Graduate Medical Education (ACGME). Surgery Milestones: The Accreditation Council for Graduate Medical Education [Internet]. 2019 [Cited 4 Apr 2019]. Available from: https://www.acgme.org/globalassets/pdfs/milestones/surgerymilestones.pdf.
3. Konsil Kedokteran Indonesia. Peraturan Konsil Kedokteran Indonesia No. 73 Tahun 2020 tentang Standar Pendidikan Profesi Dokter Spesialis Bedah [Internet]. 2020 [Cited 2 Jan 2020]. Available from: https://kki.go.id/uploads/cms_file/Peraturan_KKI_No._73_Tahun_2020_tentang_Std_Pendk_Profesi_Dokter_Spesialis_Bedah-_.pdf.
4. Soeselo DA, Werdhani RA, Lalisang TJM, Purwadianto A, Harahap WA, Yolanda R, et al. Understanding surgeon professionalism in Indonesia: A qualitative study in a multicultural and resource-limited context. J Adv Med Educ Prof. 2025;13(4):294–302.
5. ABIM Foundation, ACP-ASIM Foundation, European Federation of Internal Medicine. Medical professionalism in the new millennium: a physician charter. Ann Intern Med. 2002;136(3):243–6.
6. ACS Task Force on Professionalism. Code of professional conduct. J Am Coll Surg. 2004;199(5):734–5.
7. Marshall S, Nataraja R. Chapter 29: Patient safety and surgical education. In: Advancing Surgical Education: Theory, Evidence, and Practice. Singapore: Springer; 2019. p. 327–37.
8. Alexis DA, Kearney MD, Williams JC, Xu C, Higginbotham EJ, Aysola J. Assessment of perceptions of professionalism among faculty, trainees, staff, and students in a large university-based health system. JAMA Netw Open. 2020;3(11):e2021452.
9. Davis CH, DiLalla GA, Perati SR, Lee JS, Oropallo AR, Reyna CR, et al. A review of professionalism in surgery. InBaylor University Medical Center Proceedings. Taylor & Francis. 2024;37(6):993-7.
10. Al Ansari A, Al Khalifa K, Al Azzawi M, Al Amer R, Al Sharqi D, Al-Mansoor A, et al. Cross-cultural challenges for assessing medical professionalism among clerkship physicians in a Middle Eastern country (Bahrain): Feasibility and psychometric properties of multisource feedback. Adv Med Educ Pract. 2015;6:509–15.
11. Cullen MJ, Zhang C, Sackett PR, Thakker K, Young JQ. Can a situational judgment test identify trainees at risk of professionalism issues? A multi-institutional, prospective cohort study. Acad Med J Assoc Am Med Coll. 2022;97(10):1494–503.
12. Al Hashmi W, Klassen RM. Developing a situational judgement test for admission into initial teacher education in Oman: An exploratory study. Int J Sch Educ Psychol. 2020;8(sup1):187–98.
13. Goss BD, Ryan AT, Waring J, Judd T, Chiavaroli NG, O'Brien RC, et al. Beyond selection: The use of situational judgement tests in the teaching and assessment of professionalism. Acad Med J Assoc Am Med Coll. 2017;92(6):780–4.
14. Sandars J. The use of reflection in medical education: AMEE Guide No. 44. Med Teach. 2009;31(8):685–95.
15. Boutet I, Vandette MP, Valiquette-Tessier SC. Evaluating the implementation and effectiveness of reflection writing. Can J Scholarsh Teach Learn. 2017;8(1):1-18.
16. Soleimani-Nouri P, Entwistle OH, Fehervari M, Spalding D. A novel framework for surgical reflection. Annals of Laparoscopic and Endoscopic Surgery. 2023;8:1.
17. Poola VP, Suh B, Parr T, Boehler M, Han H, Mellinger J. Medical students' reflections on surgical educators' professionalism: Contextual nuances in the hidden curriculum. Am J Surg. 2021;221(2):270–6.
18. Purwadianto A, SpS S, Gunawan S, Budiningsih Y, Prawiroharjo P, Firmansyah A. Kode etik kedokteran Indonesia. Jakarta: Pengurus Besar Ikatan Dokter Indonesia; 2012.

19. Sorrel MA, Olea J, Abad FJ, de la Torre J, Aguado D, Lievens F. Validity and reliability of situational judgement test scores: A new approach based on cognitive diagnosis models. Organ Res Methods. 2016;19(3):506–32.
20. Whetzel DL, Sullivan TS, McCloy RA. Situational judgment tests: An overview of development practices and psychometric characteristics. Personnel Assessment and Decisions. 2020;6(1):1.
21. Kember D, McKay J, Sinclair K, Wong FKY. A four-category scheme for coding and assessing the level of reflection in written work. Assess Eval High Educ. 2008;33(4):369–79.
22. Gardner AK, Dunkin BJ. Evaluation of validity evidence for personality, emotional intelligence, and situational judgment tests to identify successful residents. JAMA Surg. 2018;153(5):409–16.
23. Smith KJ, Neely S, Dennis VC, Miller MM, Medina MS. Use of situational judgment tests to teach empathy, assertiveness, communication, and ethics. Am J Pharm Educ. 2022;86(6):8761.
24. Patterson F, Zibarras L, Ashworth V. Situational judgement tests in medical education and training: Research, theory and practice: AMEE Guide No. 100. Med Teach. 2016;38(1):3–17.
25. Peshkepija AN, Basson MD, Davis AT, Ali M, Haan PS, Gupta RN, et al. Perioperative self-reflection among surgical residents. Am J Surg. 2017;214(3):564–70.
26. Ganni S, Botden SMBI, Schaap DP, Verhoeven BH, Goossens RHM, Jakimowicz JJ. "Reflection-Before-Practice" improves self-assessment and end-performance in laparoscopic surgical skills training. J Surg Educ. 2018;75(2):527–33.
27. Aziz A, Mahboob U, Saleem T. Benefits of reflective writing in health care through the vivid lens of house officers. MedEdPublish. 2020;9:60.