Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Three-Year Generalizability Study of Clinical Skills Assessment: Improvement in Reliability Following Expansion of Clinical Reasoning Tasks

Soo-Hyun Lee
Pusan National University
Seong-Eun Kwon
Pusan National University Korean Medicine Hospital
Man-Suk Hwang
Pusan National University
Minna Hong
Pusan National University
Eui-Hyoung Hwang
Pusan National University
Jihwan Kim
Pusan National University
Aram Jeong
Gachon University
Hye-Yoon Lee
drlee@pusan.ac.kr

Pusan National University

Research Article

Keywords: Generalizability Theory, Objective Structured Clinical Examination, Clinical Competency Assessment, Reliability, Complementary and Integrative Medicine, Educational Measurement, Formative Feedback

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10539676/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Original Article

A Three-Year Generalizability Study of Clinical Skills Assessment: Improvement in Reliability Following Expansion of Clinical Reasoning Tasks

Soo-Hyun Lee[1], Seong-Eun Kwon[2], Man-Suk Hwang[3], Minna Hong[4], Eui-Hyoung Hwang[3]; Jihwan Kim[5]; Aram Jeong[6*], Hye-Yoon Lee[4,7*]

[1]School of Korean Medicine, Pusan National University, Yangsan, Republic of Korea

[2]Pusan National University Korean Medicine Hospital, Yangsan, Republic of Korea

[3]Department of Korean Medicine Rehabilitation, 3rd division of clinical medicine, School of Korean Medicine, Pusan National University, Yangsan, Republic of Korea

[4]Center for Korean Medicine Education, Research Institute for Korean Medicine, Pusan National University, Yangsan, Republic of Korea

[5]Department of Sasang Constitutional Medicine, 4th division of clinical medicine, School of Korean Medicine, Pusan National University, Yangsan, Republic of Korea

[6]Department of Pediatrics, College of Korean Medicine, Gachon University, Seongnam, Republic of Korea

[7]Division of Humanities and Social Medicine, School of Korean Medicine, Pusan National University, Yangsan, Republic of Korea

***Co-Corresponding authors**

Hye-Yoon Lee

Email: drlee@pusan.ac.kr

Tel: +82-10-9476-1836

Address: 50612, 317 School of Korean Medicine, Pusan National University, Geumo-ro, Mulgeam-eup, Yangsan-si, Gyengsangnam-do, South Korea

Aram Jeong

Email: youi22@gachon.ac.kr

Tel: +82-032-770-1292

Address: 13120, 1342 Seongnam-daero, Sujeong-gu, Seongnam-si, Gyeonggi-do 13120, South Korea

Abstract

Background: As competency-based education has expanded among health professionals, greater emphasis has been placed on teaching and assessing clinical skills as a domain distinct from knowledge. Given that clinical skills assessments involve multiple facets, establishing their validity and reliability is crucial. This study evaluated the validity and reliability of a summative clinical skill assessment (S-CSA) at **a single college over three consecutive years** (2023–2025).

Methods: The S-CSA consists of clinical performance examinations (CPX) using patient-encounter cases and objective structured clinical examinations (OSCE) using clinical skill cases. Data from three years were analyzed, with 47, 46, and 56 examinees in 2023, 2024, and 2025, respectively. The analyses included descriptive statistics, correlations among CPX domain scores, CPX total scores, and OSCE total scores, and estimation of generalizability coefficients using generalizability theory (G-theory).

Results: Descriptive analyses showed a consistent tendency for CPX scores to be lower than OSCE scores. In particular, the history-taking and physical examination domains tended to show lower achievement levels across the years. Correlations among CPX domains were modest to moderate, whereas correlations between OSCE and CPX total scores remained low, suggesting that the CPX and OSCE assessed distinct constructs. The G-coefficients ranged from 0.534 to 0.850. Reliability estimates varied across cohorts but became more stable over time, with fewer poorly functioning stations identified in later years. Sensitivity analyses demonstrated that the exclusion of underperforming stations substantially improved reliability estimates in earlier cohorts.

Conclusions: Over three years, repeated analyses of the reliability of the clinical skills assessment showed that the examination ultimately reached a level generally considered

acceptable for high-stakes assessment. This improvement may have been supported by the identification of stations that undermined the overall reliability and by the increasing number of CPX stations emphasizing clinical reasoning. Continued case development and further refinement are required. Future studies should include item-level reviews and analyses with larger sample sizes.

Keywords: Generalizability Theory, Objective Structured Clinical Examination, Clinical Competency Assessment, Reliability, **Complementary and Integrative Medicine,** Educational Measurement, Formative Feedback

Background

As competency-based education has expanded among health professionals, greater emphasis has been placed on teaching and assessing clinical skills as a domain distinct from knowledge [1, 2]. **In parallel, core competencies for primary care have been articulated, and authentic assessment formats, such as encounters with standardized patients (SPs) or simulators, are increasingly used** to evaluate students' ability to perform clinical reasoning and procedural tasks in realistic settings [3, 4].

Objective structured clinical examinations (OSCEs) have been widely used in this context. They accurately score each step of clinical skills using predeveloped checklists. Some studies have reported the Clinical Performance Examination (CPX) separately, which multidimensionally assesses students' abilities in history-taking (HT), physical examination (PE), patient education (ED), and patient–physician interaction (PPI) in patient-encounter cases [2, 3, 5]. Clinical skills assessment (CSA) encompasses the two forms of examination.

CSA faces challenges in ensuring the reliability of its evaluation results, as it involves multiple facets, such as raters, case specificity, consistency of the standardized patient's (SPs) acting, and environmental conditions on the day of the exam [6]. As CSA has become a high-stakes test linked to grade failure, graduation assessment, and national examinations, ensuring that evaluation results accurately reflect students’ actual competence and remain consistent has become the responsibility of educational institutions [7-10].

South Korea has a dualized medical system consisting of medical doctors (MDs) and doctors of Korean medicine (KMDs), which is a traditional, integrative, and complementary medicine (TICM). Previous research in the field of TICM has focused on the practical application and diversification of assessment tools [11] or has reported experiences of implementation [12, 13]. However, beyond the introduction of individual modules, few studies have evaluated the

quality of summative clinical skill assessments (S-CSAs) and provided valid evidence to support the interpretation of the assessment results. To address this gap, examining the validity of S-CSA using established assessment frameworks is necessary. While substantial evidence has accumulated in conventional medical education, similar investigations remain limited in Korean medicine education.

Validity refers to the degree to which evidence and theory support the interpretation of assessment scores for their intended use. According to Messick's framework, evidence of validity can be gathered from multiple sources, including content, response processes, internal structure, relationships with other variables, and consequences of assessment decisions [11]. Among these sources, evidence based on the internal structure is particularly important in high-stakes clinical skills assessments because it reflects the consistency and reliability of examination scores [1, 2, 12].

Generalizability Theory (G-theory) offers a robust statistical framework for evaluating the internal structure by decomposing total score variance into multiple sources of measurement error. Previous studies have used G-theory to examine the influence of factors such as student ability, item difficulty, evaluator stringency, and their interactions on assessment reliability [2, 16]. A useful feature of G-theory is its ability to identify specific sources of measurement error and detect cases or items that may undermine reliability. Furthermore, through a decision study (D-study), expected changes in reliability can be estimated under alternative assessment designs, such as varying the number of stations or evaluators, thereby supporting the development of efficient and cost-effective assessment systems [3, 17, 18].

Therefore, this study evaluated the validity of the S-CSA using evidence from the content, response process, relationships among assessment components, and internal structure. Particular emphasis was placed on reliability analysis using G-theory. By conducting repeated

evaluations over three consecutive years, we examined whether the reliability of the assessment, represented by G-coefficients, improved and became more stable over time, while identifying potential sources of measurement error and directions for quality improvement.

Methods

The structure and process of the summative clinical skills assessment (S-CSA)

The S-CSA consists of CPXs and OSCEs. CPX is defined as a simulated medical encounter with SPs operating on pre-developed clinical cases and standardized guidelines. In this assessment, examinees independently lead clinical history-taking and physical examination without a predetermined script, as in actual clinical practice. Evaluators then assess whether essential clinical reasoning and tasks are logically and accurately performed across four domains: HT, PE, ED (collectively termed non-PPI domains), and PPI. OSCEs are defined as a combination of two to three relevant CSA tasks and mainly evaluate the accuracy of procedures.

The S-CSA was administered annually to fourth-year students over two consecutive days, with two examination sessions conducted per day. Students were randomly allocated to one of the four sessions and progressed through a series of stations in a predetermined sequence, completing either the CPX or an OSCE at each station within 12 min. Faculty raters stationed at each site directly observed and assessed students' performance in the non-PPI domains of the CPX and all OSCE checklist items, whereas PPI was assessed by SPs.

The overall item composition of the CSA has been progressively revised each year to enhance assessment quality and balance the examination stations. Specifically, the configuration was four CPX and four OSCE items in 2023, six CPX and three OSCE items in 2024, and eight CPX and one OSCE item in 2025.

Data Collection

This study retrospectively collected data after the completion of the S-CSA. The S-CSA has been a regular student assessment tool for over 10 years, and this study analyzed data from the last 3 years by obtaining the results anonymously. This study was approved and granted exemption by our institutional review board (IRB approval No. GIRB-25-112).

Analysis Methods

Content validity

Content validity was considered based on the level and necessity of the content. This was considered relevant when the cases matched the blueprints.

Response Process

The response process was used to determine the minimum risk of error during the overall procedure. The directions for SPs, faculty rater instructions, and whether the CSA stems were written following the general rules for student assessments were investigated.

Internal Structure

(1) Descriptive analysis

For each case, the mean, standard deviation, median, maximum, and minimum scores were reported. Although CPX items are typically developed with weights of HT 40%, PE 30%, ED 10%, and PPI 20%, item developers were allowed to adjust these weights according to the relative importance of each domain within a given case. To enable comparison across cases and domains, scores in this paper are reported on a 100-point achievement scale rather than as raw weighted totals.

All scores are presented after rescaling to a 100-point maximum. For example, where scores are presented by case, each case was rescaled to 100; where scores are presented by domain within a case, the maximum of each domain was rescaled to 100, and the

achievement percentage is reported.

When comparing CPX and OSCE, achievement was reported by rescaling the sum of all CPX scores to 100 and the sum of all OSCE scores to 100, respectively. CSA achievement was likewise reported each year by rescaling the total practical examination score to 100.

The scores reflected area-specific proportions. For comparison, the scores for each case (by station), area-specific CPX scores, total CPX scores (four items in 2023 and six items in 2024), total OSCE scores (four items in 2023 and three items in 2024), and total practical examination scores (eight items in 2023 and nine items in 2024) were converted to a 100-point scale and reported.

(2) Inter-area correlation analysis

Correlation analyses were performed among the CPX domains, CPX total, OSCE, and practical total scores. Pearson's correlation test and IBM SPSS version 22.0 (IBM Corp., Armonk, NY, USA) were used.

(3) Reliability

The reliability of the practical performance score was evaluated using G-theory, which is a method for calculating reliability when various factors influence the score. This analysis focused on two facets that affect practical performance scores: case and item. The design employed was P×(I:C), where case (C) was crossed with student (person, P) and item (I) was nested within C. The analysis identified the influence of each factor and their interactions, namely P, C, I:C (nested within C), PC, and PI on the total variance of the evaluation score, expressed as a percentage (variance due to each variable/total variance). Here, P represents differences in student ability, C indicates the difficulty level of the case, I:C denotes the difficulty level of the item within the case, PC reflects differences between students across

cases (case specificity), and PI:C accounts for the remaining unidentifiable errors. The reliability of the evaluation was estimated using variance components derived from G-theory, and generalizability (G) and phi coefficients were reported. The G String IV program, version 6.3.2 (Bloch & Norman, Hamilton, ON, Canada), was used for the analysis because it is optimized for decomposing total score variance into various facets, such as student ability and item difficulty.

Generalizability theory was employed as the primary analytical framework because it provides a robust method for calculating reliability when multiple complex error factors influence performance scores [14]. Particularly in the context of OSCEs and clinical performance assessments, G-theory allows researchers to decompose observed score variance into fine-grained categories and simultaneously assess the magnitude of error arising from multiple facets, such as raters, cases, and standardized patients [15]. Furthermore, it facilitates decision studies (D-studies) that help optimize the assessment design by simulating the impact of adjusting these facets, thereby ensuring the defensibility and consistency of high-stakes clinical evaluations [16].

Results

Content Validity

Practical cases were selected and developed based on the examination blueprint. The initial draft of each case was developed by experts in the relevant medical fields. Case selection was based on pre-identified common clinical presentations, and the cases were subsequently revised through review by an expert group specializing in medical education and assessment before being finalized. Following case development, SPs were trained, and rehearsals were conducted to ensure accurate and standardized case portrayals.

The clinical presentations were first identified, after which content experts developed each item at a level appropriate for the fourth-year medical students, using cases relevant to primary care. After item development, the lead faculty member, who had 5 years of experience managing the comprehensive examination and 15 years of clinical experience, together with the team-teaching faculty, reviewed each item to ensure its clinical relevance, educational appropriateness, and consistency with the examination blueprint.

Response Process

The following procedures were designed to support response process validity by reducing rater inconsistency and ambiguity in the scoring criteria.

Rater preparation

A faculty rater was assigned to each station to directly observe and assess students. The raters reviewed the assessment items in advance and had the opportunity to ask questions and discuss their interpretations before the examination. Faculty evaluators underwent training on PE items before the examination and were provided with reference videos. On the day of the examination, the faculty raters were instructed to familiarize themselves with the scoring criteria before the exam began.

Scoring tool design

The checklists were developed using a binary 1-0 scoring scheme based on whether each task was performed. The items used plain terminology accessible to non-specialists, and each checklist item was constrained to assess a single behavior, thereby minimizing ambiguity in scoring decisions. For PE items, supplementary videos or photographic materials were provided to support the consistent application of the criteria.

Standardized patient training

Among the CPX items, PPI was scored directly by SPs. A clinical professor led the SPs' training. During a 4-h pre-training pre-examination training session, SPs viewed recorded CPX encounters, scored them, and discussed their judgments with other evaluators to promote consistent application of the PPI scoring criteria. The SPs underwent repeated mock-scoring sessions to standardize their rating criteria before the actual examination.

Internal Structure

Descriptive statistics

During the first examination session (Group A) of the 2023 academic year, the shoulder PE station achieved the highest mean score (90.6 ± 10.9), whereas the cancer-related fatigue achieved the lowest mean score (67.5 ± 11.7). During the second examination session (Group B), the sensory examination and splinting achieved the highest mean score (92.8 ± 9.4), whereas the hypertension wound dressing station achieved the lowest average score (62.8 ± 11.6). Detailed analyses indicated relatively low achievement in the PE and HT domains. Overall, the OSCE stations produced higher scores than the CPX stations. (Table 1)

During the first examination session (Group C) pf the 2024 academic year, the adult basic cardiopulmonary resuscitation and automated external defibrillation station and the wound dressing station achieved the highest mean score (88.2 ± 7.6), whereas the headache station scored the lowest mean score (69.8 ± 8.5). During the second examination session (Group D),

the adult basic cardiopulmonary resuscitation and automated external defibrillation station and the wound dressing station again achieved the highest mean score (91.1 ± 7.0), whereas the temporomandibular joint pain station achieved the lowest average score (71.8 ± 7.1). The PPI and medical history-taking domains showed relatively lower scores than the remaining domains. Overall, the OSCE stations produced higher scores than the CPX stations. (Table 2)

During the first examination session (Group E) of the 2025 academic year, the chronic cough station achieved the highest mean score (89.1 ± 6.44), whereas the allergic rhinitis station scored the lowest mean score (48.5 ± 6.63). During the second examination session (Group F), the chronic cough station again achieved the highest mean score (87.0 ± 7.37), whereas the hand tremor station achieved the lowest average score (54.2 ± 8.80). Detailed analyses indicated relatively low achievement in the PE domains in Group E, whereas PE showed higher achievement in Group F. Overall, the OSCE stations produced higher scores than the CPX stations. (Table 3)

Correlation analysis by domain

The degree of correlation was interpreted according to Dancey and Reidy's criteria: $|r|<0.1$ indicates none, $0.1\leq|r|<0.4$ indicates weak, $0.4\leq|r|<0.7$ indicates moderate, $0.7\leq|r|<1$ indicates strong, and $|r|=1$ indicates perfect agreement. Correlations among the CPX domains, CPX total scores, and OSCE scores were then examined. A statistically significant correlation was observed only between the PPI and OSCE scores.

For the 2023 academic year, CPX total scores showed strong correlations with HT, PE, and ED (r = 0.835, 0.772, and 0.732, respectively) and a moderate correlation with PPI (r = 0.569). PPI showed a moderate correlation with PE (r = 0.503). The correlation between CPX and OSCE scores was weak (0.140).

For the 2024 academic year, CPX total scores showed strong correlations with HT, PE, and ED (r = 0.698, 0.712, and 0.787, respectively) and a weak correlation with PPI (r = 0.569). The correlation between CPX and OSCE was weak (r = 0.211).

For the 2025 academic year, CPX total scores showed strong correlations with HT, PE, and ED (r = 0.757, 0.769, and 0.621, respectively). The correlation with PPI was weak (r = 0.228) and was not statistically significant. Correlations between PE and ED (r = 0.341), PE and PPI (r = 0.218), and ED and PPI (r = 0.175) ranged from weak to moderate. The correlation between CPX and OSCE was weak (0.144) and was not statistically significant. Overall, CPX total scores were more strongly correlated with objective clinical performance assessment domains (HT, PE, and ED) than with PPI and OSCE scores. (Table 4)

Reliability Analysis

The reliability analysis results were expressed as the percentage of variance explained, with variance components (VCs) reported as percentages of the total variance.

(1) Reliability of the 2023 assessment

In the non-PPI domain for Group A, the variance due to items nested within the case (I:C), excluding the unidentifiable error (PI:C), was the highest at 17.39%. The variance due to case-specific differences among students (PC; case specificity) was 5.43%, and the variance due to students (P; student discrimination) was 1.09%. The generalizability G coefficient was 0.569. For Group B, the variance due to items (I:C), excluding the PI:C, was the highest at 21.08%, with the variance due to case-specific differences among students (PC) at 2.94% and the variance due to students (P) at 3.92%. The G coefficient was 0.850 (Table 5).

In the PPI domain, Group A showed the highest variance attributable to the person-by-item interaction (PI; item specificity) at 77.21%, followed by the variance due to items

(I; item discrimination) at 15.61% and persons (P; person discrimination) at 7.18%. The G coefficient was 0.848. For Group B, the person-by-item interaction (PI) similarly accounted for the largest proportion of variance at 73.21%, followed by items (I) at 23.17% and persons (P) at 3.62%. The G coefficient was 0.748 (Table 6).

(2) Reliability of the 2024 assessment

For non-PPI items in Group C, the variance due to items nested within cases (I:C), excluding the unidentifiable error (PI:C), was the highest at 35.79%. The variance attributable to case-specific differences among students (PC; case specificity) was 1.03%, and the variance due to students (P; student discrimination) was 0.70%. The generalizability G coefficient was 0.534.

In Group D, the variance of items nested within cases (I:C), excluding the unidentifiable error (PI:C), was the highest at 21.30%; the variance attributable to case-specific differences among students (PC) was 7.21%, and the variance due to students (P) was 3.40%. The generalizability G coefficient was 0.568 (Table 5).

For the PPI items in Group C, the variance due to items (I; item discrimination) was the highest at 53.96%, the variance due to persons (P; person discrimination) was 4.34%, and the person-item interaction (PI; item specificity) was 41.69%. The generalizability of the G coefficient is 0.789. In Group D, the person-item interaction (PI; item specificity) was highest at 57.97%, the variance due to persons (P; person discrimination) was 4.57%, and the variance due to items (I; item discrimination) was 37.46%. The generalizability of the G coefficient was 0.632 (Table 6).

(3) Reliability of the 2025 assessment

For the non-PPI domain in Group E, excluding the unidentifiable error (PI:C), the variance

due to items nested within cases (I:C) was the highest at 58.11%, whereas the variance attributable to case-specific differences among students (PC; case specificity) was 0.05% and the variance due to students (P; student discrimination) was 0.81%. The generalizability of the G coefficient was 0.805.

In Group F, the variance due to items (I:C) remained at 58.75%, the case-specific variance of students (PC) was 0.70%, and the variance due to students (P) was 0.60%. The generalizability of the G coefficient is 0.692 (Table 5).

In the PPI domain for Group E, student-by-item discrepancy (PI; item specificity) accounted for the largest proportion of variance (72.22%), followed by variation due to students (P; student discrimination) at 17.56% and variation due to items (I; item discrimination) at 10.22%. The generalizability of the G coefficient was 0.593. In Group F, the student-by-item discrepancy (PI) accounted for the largest proportion of variance at 69.89%, followed by variance due to students (P) at 21.12% and variance due to items (I) at 8.99%. The generalizability of the G coefficient was 0.645. (Table 6)

(4) Sensitivity analysis: reliability after exclusion of underperforming stations

To examine how individual stations contributed to overall reliability, we conducted a sensitivity analysis in which stations with notably low mean scores or atypical score distributions were excluded and the G-study was repeated. Specifically, in 2023, the OSCE pulmonary function test, transfer rate, and MMSE items were excluded from Group A; in 2024, the OSCE acupuncture therapy, Chuna, and splinting items were excluded from Group C, and the CPX headache item was excluded from both Groups C and D. No stations were excluded for the 2025 cohort, as all stations contributed comparably to score variance.

After exclusion, the non-PPI G-coefficient for the 2023 Group A increased from 0.569 to 0.682, while Group B remained unchanged at 0.850. For the 2024 cohort, the G-coefficient for Group C rose from 0.534 to 0.667, and for Group D from 0.568 to 0.610. Phi coefficients showed a similar pattern of improvement (e.g., 2023 Group A: 0.434 to 0.501; 2024 Group C: 0.450 to 0.541). Detailed results are presented in Supplementary data 1.

These changes suggest that a small number of poorly functioning stations had a substantial influence on the overall reliability estimates in the first two years and that station-level review can yield meaningful gains in reliability without altering the overall assessment design. (Supplementary data 1)

Discussion

This study evaluated the validity of summative clinical skills assessment (S-CSA) over a three-year period by examining the evidence related to content, assessment procedures, and internal structure based on Messick's framework.

Three years of CSA data yielded several main observations. First, OSCE scores were consistently higher than CPX scores, reflecting the greater cognitive demand of integrated case-based assessment compared with discrete procedural skills. Second, performance varied substantially by domain: PE scores improved over time alongside iterative faculty feedback, whereas PPI scores remained consistently low. Third, the non-PPI G coefficient reached 0.805–0.850 by the third year, meeting conventional thresholds for high-stakes assessments, with variance attributable to items nested within cases accounting for a greater proportion of score variance than case-specific variance. Fourth, the weak correlations between CPX and OSCE total scores (0.140–0.211) support their treatment as distinct constructs. The remainder of the discussion examines the implications of these findings for the design and interpretation of clinical skills assessment.

Reliability adequate for high-stakes use

The G-study demonstrated that, among all groups, Group B exhibited the highest G coefficient (0.850) and phi coefficient (0.731) among all groups in the non-PPI area for the 2023 academic year. A G coefficient of 0.8 is commonly regarded as an indicator of sufficient reliability (with 0.7 a frequently used one), although the appropriate threshold depends on the stakes of the assessment [17]. Therefore, the value of 0.850 observed here exceeds commonly accepted standards for high-stakes decisions such as student retention or graduation **[18, 19]**. This finding indicates that the assessment produced sufficiently consistent scores to support consequential academic decisions while minimizing measurement error. Accordingly, these

results provide evidence supporting the reliability component of the validity argument for the S-CSA **[15, 20, 21]**.

Sources of score variance: items within cases, not case specificity

A major advantage of G-theory is its ability to identify sources of score variance [22]. Although case specificity has traditionally been regarded as a major threat to OSCE reliability [23, 24], variance attributable to case-specific differences among students accounted for only 2.94% of total score variance in this study, whereas items nested within cases accounted for 21.08% [25]. This finding suggests that score differences were influenced less by variation in case difficulty than by the characteristics of the checklist items within each case.[26, 27] Beyond the relative magnitudes, the substantial contribution of I:C indicates that the items exhibited appropriate difficulty and discrimination, with student performance well differentiated by item quality. These findings underscore the importance of careful item construction, difficulty calibration, and item selection during assessment development to maximize reliability.

The sensitivity analysis showed that overall reliability increased after exclusion of stations that were suspected of reducing assessment, suggesting that individual stations differed in their contribution to score consistency. Existing checklists often emphasize a comprehensive procedural approach by verifying whether a student has performed all textbook procedures without omissions. However, simply enumerating procedural steps may reduce the assessment's ability to distinguish students [28] with highly developed clinical skills from those with weaker clinical reasoning abilities, because students may demonstrate thorough data collection while struggling with clinical reasoning [29]. A previous study revealed that the assessment accuracy improved when the checklist was reorganized around key actions that influenced diagnosis and treatment, specifically those that were clinically discriminatory [30].

Therefore, future item development should avoid a simple enumeration of actions and instead focus on adjusting the difficulty and weighting of items centered on the core indicators necessary for the clinical reasoning process.

Traditional OSCEs assessing modality-specific techniques and procedural skills typically require approximately 5 minutes per station. Increasingly, these assessments have evolved into standardized patient (SP)-encounter-based OSCEs to enhance clinical authenticity and better simulate real-world patient encounters. Although this transition may improve assessment authenticity and standardization, recent studies emphasize that developing robust assessment methods integrating procedural skills with clinical reasoning, together with establishing appropriate station-specific passing standards, remains essential for ensuring the quality of clinical skill examinations [31].

Domain-level performance patterns

When CPX performance was examined by domain, patient education (ED) showed the highest scores, whereas physical examination (PE) and patient–physician interaction (PPI) showed relatively lower scores.

Lower PE scores have been related to insufficient practice time. Although this study did not directly investigate the reasons for the lower scores, further evaluation is needed to determine whether students receive adequate opportunities and time to learn and practice PE skills and whether instruction sufficiently addresses not only the performance of PE itself but also the clinical decision-making involved in selecting appropriate examination procedures. Despite the initially lower scores, PE performance improved steadily from 71.0 ± 8.1 in 2023 to 81.5 ± 6.5 in 2024 and 84.6 ± 7.2 in 2025. This trend may reflect the cumulative effects of faculty feedback and the increased emphasis on PE training in clinical practice settings. These findings reinforce the importance of PE as a core clinical competency and highlight the need for

continued opportunities for deliberate practice and feedback.

In contrast, ED scores, although generally high, decreased slightly in 2025. Because patient education requires clinical reasoning, including explanations of diagnostic plans, presumptive diagnoses, safety-netting, and individualized management strategies, this decline may reflect increasingly demanding assessment criteria rather than reduced student competence. Further refinement of this domain, including the incorporation of patient notes or structured reasoning assessments, may facilitate a more precise evaluation of clinical reasoning.

PPI scores remained consistently low. In particular, the item "Structuring the Interview" appeared especially challenging. Unlike history-taking and PE items, which are largely scored using dichotomous checklists, PPI relies on a four-point rating scale assessing qualitative aspects of performance, creating a higher threshold for top scores. Future refinement of PPI assessment may require modification of the rating scale and greater emphasis on educational interventions aimed at developing patient-centered communication skills. **Additionally, case-level score distributions ranged from hypertension (62.8) to dyspepsia (DYS; 89.6) in 2023; from headache (69.8) to cough (82.0) in 2024; and from acute rhinitis (AR; 48.5) to chronic cough (89.1) in 2025. Students performed particularly well in musculoskeletal pain cases, such as shoulder PE (90.6) and low back pain (83.4). This pattern may reflect content-specific performance [32], whereby musculoskeletal cases permit more direct observation of procedural skills than internal medicine cases, such as hypertension (62.8) and cancer-related fatigue (67.5) [33]. Additionally, students may have benefited from greater clinical exposure and repeated practice with musculoskeletal conditions during clinical training [34], leading to better internalization of these skills [35].**

CPX and OSCE as distinct assessment constructs

Correlation analysis showed that CPX total scores were strongly associated with history

taking, physical examination, and patient education, whereas correlations with PPI were weaker. This finding suggests that the CPX captures multiple dimensions of competence and that communication skills function as a distinct construct from information gathering and clinical reasoning. The relatively low PPI scores and weaker correlations further support the view that communication competency represents a noncognitive attribute requiring targeted educational interventions, consistent with findings from studies of the USMLE Step 2 CS [36].

The correlation between the CPX and OSCE total scores was low in both 2023 and 2024, as well as in 2025 ($r = 0.140$, 0.211, and 0.144, respectively), providing evidence that the two assessments measured different dimensions of competence. The OSCE primarily evaluates standardized procedural skills that can be repeatedly practiced and assessed using structured checklists, [37] whereas the CPX requires context-specific performance across diverse clinical presentations, reflecting the complexity of clinical competence. Thus, whereas the OSCE focuses on technical proficiency, the CPX places greater emphasis on clinical reasoning and integrated decision-making.

Although the OSCE primarily evaluates technical proficiency acquired through repeated practice and standardized checklists, the CPX serves as a higher-order assessment of clinical reasoning by requiring students to integrate patient information, generate differential diagnoses, and make context-dependent clinical decisions. Furthermore, previous studies have demonstrated that CPX reliability is sensitive to the number of cases administered. Specifically, at least three cases are required to achieve a G coefficient of 0.70 or higher [38] whereas more than six cases are generally needed to achieve a G coefficient of 0.80 [38]. The improved reliability observed following the expansion of the number of CPX stations in the present study supports these findings and suggests that increasing the number of clinically relevant cases enhances the discriminative capacity and reliability of the assessment.

Differential domain weighting in composite scoring

Because the CPX and OSCE measure largely independent competencies [39, 40], students who perform well in technical procedures may not necessarily demonstrate comparable proficiency in patient history-taking or patient-physician interaction.

Accordingly, combining scores from all domains into a single composite score without considering differences in the constructs being measured may reduce the precision of the overall assessment [41]. Prior studies applying G-theory suggest that overall score error can be minimized by assigning greater weight to domains with higher reliability estimates (i.e., domains with higher generalizability) [29, 36]. This study confirmed differences in reliability across domains; therefore, future evaluation design should consider applying differential weighting to each domain based on its psychometric characteristics [42, 43]. This method aims to minimize distortion when integrating complex clinical competencies into a single score.

Assessing PPI and other non-cognitive domains

The lowest PPI reliability estimate was observed in Group D (G = 0.632) [44]. This finding suggests that non-cognitive competencies, such as professional attitude and communication, are inherently more difficult to evaluate objectively [45], leaving greater scope for evaluator subjectivity [32]. In particular, the large person-by-item interaction variance (PI) observed in this study indicates that student performance varied substantially across individual rating items, reflecting inconsistency in how students performed on different aspects of PPI rather than differences among evaluators alone [46].

Two approaches may help address this issue. First, evaluator training should be strengthened. To minimize subjective differences between evaluators, structured calibration sessions involving shared scoring criteria, discussion through exemplar performances, and mock-scoring exercises should be conducted before the examination [23, 47]. Second, greater use of

a Global Rating Scale should be considered. Previous research suggests that, when assessing non-cognitive competencies, global judgments made by experienced evaluators may provide more reliable assessments than detailed behavioral checklists alone [48]. Therefore, combining structured checklists with global rating scales may improve the assessment of PPI and other communication-related competencies [49]

Consistent with this interpretation, the PPI domain in Group D during the 2024 academic year demonstrated the lowest G coefficient (0.632), although its phi coefficient remained relatively high (0.739), indicating differences between relative and absolute reliability. The person-by-item interaction (PI) accounted for 57.97% of the total variance, suggesting considerable variation in student performance across individual PPI items. Additionally, item variance (I) accounted for 37.46% of the total variance, indicating significant differences in item difficulty. By contrast, variance attributable to students (P) was only 4.57%, indicating the differences among students contributed relatively little to the overall score variance. One possible explanation is that the PPI items differed considerably in their cognitive and communication demands. For example, the items assessing patient care during the PE primarily evaluated behaviors such as hand hygiene and communication regarding patient discomfort, which students generally performed well on. In contrast, the item assessing whether the interview was conducted in a logical and structured manner required students to integrate communication skills, clinical reasoning, history-taking, physical examination planning, and patient management simultaneously, making it substantially more demanding.

Limitations

This study has some limitations. First, the study was conducted at a single institution with approximately 50 students per year, which may have limited the generalizability of the findings. Second, some checklist items may contain overlapping content, potentially reducing the

discriminatory power. Despite these limitations, this study provided valuable evidence regarding the validity and reliability of summative clinical skills assessments in Korean medicine education.

Future studies should include multi-institutional data and larger sample sizes to enhance generalizability. Given the complexity of clinical reasoning, additional assessment approaches such as Global Rating Scales and repeated assessments across multiple time points should be considered to complement checklist-based evaluations and provide a more comprehensive understanding of student development.

Conclusion

This study compared and analyzed the results of clinical skills examinations (OSCE and CPX) for the academic years 2023, 2024, and 2025, confirming that although students demonstrated high performance in essential technical skills, improvements in PPI and complex clinical reasoning are needed. Excluding items with low average scores to enhance evaluation reliability underscores the need for continuous improvement of evaluation tools. The OSCE effectively assesses basic technical skills, whereas the CPX evaluates integrated problem-solving abilities in various clinical situations.

Beyond merely assessing checklist item fulfillment, accurately measuring students' high-order thinking skills, particularly their clinical reasoning processes, is crucial. Clinical reasoning is a complex concept with various definitions, necessitating diverse assessment methods for a comprehensive evaluation. A 'clinically discriminating checklist' that emphasizes items that significantly impact diagnostic decision-making demonstrates higher accuracy and psychometric stability than a comprehensive checklist. Given the complexity of clinical reasoning, fully assessing student competency using a single method is challenging. Therefore, a multifaceted assessment approach that considers the strengths and weaknesses of each method is essential to cultivate effective healthcare professionals. Ultimately, this multimethod approach will enhance the measurement of reasoning skills, improve the quality of medical education, and help develop competent healthcare personnel to ensure patient safety.

List of abbreviations

AED, Automated External Defibrillator; AMN, Amenorrhea; AR, Allergic Rhinitis; CMT, Chuna Manual Therapy; CPR, Cardiopulmonary Resuscitation; CPX, Clinical Performance Examination; CRF, Cancer-related Fatigue; DYS, Dyspepsia; ED, Patient Education; HT, History-taking; MAT, Manual Acupuncture; MMSE, Mini-Mental State Examination; MPP, Meridian Point Pressure Test; MUC, Male Urethral Catheterization; OSCE, Objective Structured Clinical Examination; PE, Physical Examination; PFT, Pulmonary Function Test; Pharm, Pharmacopuncture; PPI, Patient-Physician Interaction; SE, Sensory Examination; SP, Splinting; TMJP, Temporomandibular Joint Pain; TOS, Thoracic Outlet Syndrome; VBS, Venous Blood Sampling; VP, Venipuncture.

Declarations

Ethics approval and consent to participate

This study was approved and granted an exemption by the Institutional Review Board of Gachon University Gil Oriental Medical Hospital (approval No. GIRB-25-112) and conducted in accordance with the Declaration of Helsinki. Consent to participate was not applicable as this was a retrospective study using anonymised administrative examination data.

Consent for publication

Not applicable.

Availability of data and materials

The datasets used and/or analyzed in the current study are available from the corresponding author upon reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was supported by the Korea Institute of Korean Medicine Education and Evaluation (2025) and a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. RS-2026-25493305).

Authors' contributions

Data analysis: SHL, SEK

Study design and management: AJ, HYL

Writing - original draft: SHL

Writing - revision: MSH, MH, HYL, EH, JK

All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

Table 1. Scores of cases CPX/OSCE (n = 23 in Group A, n = 24 in Group B, 2023)

	Case	Mean ± SD	Low	High	Median	Case	Mean±SD	Low	High	Median
	Group A (n = 23)					**Group B (n = 24)**				
CPX by station	DYS	89.6 ± 9.1	69.8	98.1	92.5	DYS	67.5±12.7	43.4	86.8	68.9
	Facial Palsy	78.6 ± 8.2	62.7	96.1	80.4	Low back pain	83.4±7.4	69.0	95.2	82.1
	Tension	84.8 ± 7.5	69.4	97.2	83.3	Hypertension	62.8±11.6	44.2	90.7	62.8
	CRF	67.5 ± 11.7	47.9	87.5	68.8	CRF	80.6±7.9	62.5	95.8	81.3
CPX by domain	HT	80.6 ± 6.1	67.7	92.1	81.1	HT	72.5±8.5	52.7	86.6	72.8
	PE	81.3 ± 6.7	67.9	94.3	80.2	PE	71.0±8.1	55.6	91.1	77.8
	ED	85.3 ± 7.6	64.7	94.1	86.3	ED	76.6±4.9	68.3	86.7	78.3
	PPI	79.7 ± 6.2	72.2	92.8	78.9	PPI	78.0±5.9	56.1	82.2	75.7
	CPX total score	***79.5 ± 6.1***	***69.1***	***90.2***	***80.0***	***CPX total score***	***70.3±8.2***	***52.7***	***84.0***	***69.7***
OSCE	Emergency[#]	77.9 ± 10.6	56.3	100.0	78.1	Emergency	86.1±11.5	50.0	100.0	90.6
	PFT, MPP, MMSE	76.7 ± 7.6	62.7	90.2	76.5	VP, MUC	80.9±13.2	45.9	97.3	83.8
	Shoulder PE	90.6 ± 10.9	66.7	100.0	95.8	Musculoskeletal Diagnosis	85.9±12.7	50.0	100.0	87.5
	Pharm & cupping	78.6 ± 12.9	47.1	94.1	82.4	SE, SP	92.8±9.4	68.2	100.0	95.5
	OSCE total score	***80.1 ± 5.7***	***70.6***	***88.0***	***81.5***	***OSCE total score***	***86.7±7.1***	***68.6***	***97.6***	***87.0***

[#]Emergency Cardiopulmonary Resuscitation, Automated External Defibrillator, and Wound Dressing

DYS, dyspepsia; CRF, cancer-related fatigue; HT, history-taking; PE, physical examination; ED, patient education; PPI, patient-physician interaction; PFT, pulmonary function test; MPP, meridian point pressure test; MMSE, mini-mental state examination; Pharm, pharmacopuncture; VP, venipuncture; MUC, male urethral catheterization; SE, sensory examination; SP, splinting.

Table 2. Scores of cases CPX/OSCE (n = 18 in Group C, n = 28 in Group D, 2024)

	Case	Mean ± SD	Low	High	Median	Case	Mean ± SD	Low	High	Median
	Group C (n = 18)					**Group D (n = 28)**				
CPX by station	AMN	80.6 ± 6.4	67.6	91.2	80.9	AMN	80.3 ± 5.7	70.6	94.1	80.9
	TMJP	70.6 ± 8.7	50.0	85.7	71.4	TMJP	71.8 ± 7.1	57.1	85.7	70.2
	DYS	74.5 ± 6.6	60.4	83.3	75.0	Cough	82.0 ± 5.6	70.5	93.2	81.8
	Dizziness	74.0 ± 8.0	58.3	88.9	73.6	Neck pain	77.2 ± 7.5	62.2	91.1	76.7
	Nasal congestion	77.8 ± 6.0	68.7	88.1	77.6	Depressed mood	72.7 ± 9.1	46.9	87.8	74.5
	Headache	69.8 ± 8.5	47.9	81.3	72.9	Headache	81.0 ± 10.0	62.5	95.8	84.4
CPX by domain	HT	75.1 ± 4.2	67.7	82.8	75.8	HT	76.4 ± 7.3	60.2	90.3	75.3
	PE	67.5 ± 7.8	53.6	83.9	64.3	PE	81.5 ± 6.5	67.3	93.9	81.6
	ED	74.7 ± 9.0	47.5	85.0	77.5	ED	87.6 ± 5.8	73.3	96.7	88.3
	PPI	62.8 ± 5.0	55.6	72.2	61.1	PPI	60.6 ± 5.1	50.9	70.4	61.1
	CPX total score	***71.7 ± 4.1***	***59.9***	***78.2***	***72.3***	***CPX total score***	***77.1 ± 4.2***	***69.6***	***85.3***	***77.6***
OSCE	Emergency[#]	88.2 ± 7.6	75.0	100.0	87.5	Emergency	91.1 ± 7.0	75.0	100.0	93.8
	VP, MUC	84.3 ± 13.9	51.7	100.0	89.7	VP, MUC	81.7 ± 12.2	51.7	96.6	86.2
	Pharm., CMT, SP	81.2 ± 11.5	60.9	95.7	80.4	Shoulder PE, CMT, SP	83.4 ± 10.0	61.5	100.0	84.6
	OSCE total score	***84.5 ± 11.7***	***51.7***	***100.0***	***86.2***	***OSCE total score***	***85.4 ± 10.8***	***51.7***	***100.0***	***88.0***

[#]Emergency Cardiopulmonary Resuscitation, Automated External Defibrillator, and Wound Dressing

CPX, clinical performance examination; OSCE, objective structured clinical examination; AMN, amenorrhea; TMJP, temporomandibular joint pain; DYS, dyspepsia; HT, history-taking; PE, physical examination; ED, patient education; PPI, patient-physician interaction; VP, venipuncture; MUC, male urethral catheterization; Pharm, pharmacopuncture; CMT, chuna manual therapy; SP, splinting.

Table 3. Scores of 18 case CPX/OSCE (n = 28 in group E, n = 28 in group F, 2025)

	Case	Mean ± SD	Low	High	Median	Case	Mean±SD	Low	High	Median
	GroupE (n = 28)					**GroupF (n = 28)**				
CPX by station	Anxiety	80.4 ± 7.78	65	95	80	Facial Palsy	67.1 ± 10.07	51.5	100.0	66.7
	Knee Pain	70.9 ± 10.26	48.5	87.9	71.2	Dyspepsia	62.5 ±12.32	42.1	92.1	60.5
	Headache	66.6 ± 7.13	52.9	79.4	64.7	Low Back Pain	72.9 ± 7.28	60.0	88.0	72.0
	AR	48.5 ± 6.63	35.4	62.5	47.9	Dizziness	70.3 ± 6.92	54.6	81.8	72.7
	AMN	81.6 ± 8.13	65.0	100.0	80.0	AMN	83.4 ± 7.33	65.0	95.0	85.0
	Hand Tremor	53.3 ± 9.94	38.5	79.5	51.3	Hand Tremor	54.2 ± 8.80	35.9	71.8	55.1
	Chronic Cough	89.1 ± 6.44	73.7	100.0	89.5	Chronic Cough	87.0 ± 7.37	68.4	100.0	89.5
	TOS	68.5 ± 9.21	47.6	90.5	66.7	TOS	71.1± 7.08	61.9	85.7	71.4
CPX by domain	HT	75.1 ± 5.99	62.7	89.1	74.6	HT	73.0 ± 6.07	62.2	83.7	72.5
	PE	61.3 ± 8.04	48.1	81.0	62.03	PE	61.2 ± 6.80	49.4	76.5	59.3
	ED	75.1 ± 7.44	56.4	89.7	76.9	ED	73.0 ± 6.80	50.0	89.5	73.7
	PPI	75.8 ± 15.35	35.3	100.0	76.5	PPI	75.4 ± 15.66	37.5	100.0	76.5
	CPX total score	***69.7 ± 15.47***	***35.4***	***100.0***	***71.0***	***CPX total score***	***71.1 ± 13.10***	***35.9***	***100.0***	***71.8***
OSCE	Emergency[#]	84.6 ± 7.23	65.6	100.0	84.4	VBS, MAT	86.3 ± 10.61	62.5	100.0	87.5
	OSCE total score	***84.6 ± 7.23***	***65.6***	***100.0***	***84.4***	***OSCE total score***	***86.3 ± 10.61***	***62.5***	***100.0***	***87.5***

[#]Emergency Cardiopulmonary Resuscitation, Automated External Defibrillator, and Wound dressing

CPX, clinical performance examination; OSCE, objective structured clinical examination; AR, allergic rhinitis; AMN, amenorrhea; TOS, thoracic outlet syndrome; HT, history-taking; PE, physical examination; ED, patient education; PPI, patient–physician interaction; VBS, venous blood sampling; MAT, manual acupuncture.

Table 4. Correlation between domains of CPX/OSCE in 2023, 2024, and 2025

	2023						2024						2025					
	HT	**PE**	**ED**	**PPI**	**CPX**	**OSCE**	**HT**	**PE**	**ED**	**PPI**	**CPX**	**OSCE**	**HT**	**PE**	**ED**	**PPI**	**CPX**	**OSCE**
HT	1						1						1					
PE	0.573**	1					0.339*	1					0.342**	1				
ED	0.331*	0.468**	1				0.171	0.634**	1				0.383**	0.341*	1			
PPI	0.368**	0.503**	0.209**	1			0.093	0.134	0.036	1			0.092	0.218	0.175	1		
CPX[#]	0.835**	0.772**	0.732**	0.569**	1		0.698**	0.712**	0.787**	0.209	1		0.757**	0.769**	0.621**	0.228	1	
OSCE	0.070	0.094	0.099	0.217**	0.140	1	0.168	0.149	0.072	0.258	0.211	1	0.117	0.128	-0.106	0.020	0.144	1

HT, history-taking; PE, physical examination; ED, patient education; PPI, patient-physician interaction; CPX, clinical performance examination; OSCE, objective structured clinical examination.

[#]CPX: a composite score calculated by combining HT, PE, ED, and PPI according to a predefined ratio.

*$p < 0.05$; **$p < 0.01$

Table 5. Generalizability study of the non-PPI domains 2023, 2024 and 2025

2023							**2024**						**2025**					
Effect	Coefficient	df	SS	MS	VC	% Variance	Coefficient	df	SS	MS	VC	% Variance	Coefficient	df	SS	MS	VC	% Variance
			Group A*						**Group C**#**						**Group E**			
P	-	22	16.16	0.73	0.002	1.09	-	17	6.92	0.40	0.0015	0.70	-	27	25.120	0.93	0.0037	0.81
C	-	12	55.99	4.67	0.010	5.43	-	5	7.55	1.51	-0.0001	-0.05	-	8	136.09	17.01	0.0154	3.36
I:C	-	186	163.99	0.88	0.032	17.39	-	133	201.66	1.52	0.0767	35.79	-	188	1439.62	7.66	0.2666	58.11
PC	-	264	74.26	0.28	0.010	5.43	-	85	15.77	0.19	0.0022	1.03	-	216	42.60	0.20	0.0002	0.05
PI:C	-	4092	532.97	0.13	0.130	70.65	-	2261	303.02	0.13	0.1340	62.53	-	5,076	974.46	0.19	0.1920	41.87
Phi	0.434						0.450						0.504					
G	0.569						0.534						0.805					
			Group B						**Group D**						**Group F**			
P	-	23	38.37	1.67	0.008	3.92	-	27	23.36	0.87	0.0136	3.40	-	27	19.72	0.73	0.0025	0.60
C	-	11	61.93	5.63	0.012	5.88	-	5	73.41	14.7	0.0709	17.74	-	8	83.52	10.44	0.0055	1.30
I:C	-	171	197.78	1.16	0.043	21.08	-	30	77.54	2.58	0.0851	21.30	-	183	1,303.91	7.13	0.2481	58.75
PC	-	253	57.40	0.23	0.006	2.94	-	135	50.50	0.37	0.0288	7.21	-	216	51.82	0.24	0.0030	0.70
PI:C	-	3933	531.81	0.14	0.135	66.18	-	810	162.96	0.20	0.2012	50.35	-	4,941	877.67	0.18	0.1776	41.97
Phi	0.731						0.357						0.471					
G	0.850						0.568						0.692					

P, person; I, item; Phi, dependability coefficient; G, generalizability coefficient; df, degrees of freedom; SS, sum of squares; MS, mean square; VC, variance component.

Table 6. Generalizability study of the PPI domains 2023, 2024 and 2025

2023							**2024**						**2025**					
Effect	Coefficient	df	SS	MS	VC	% Variance	Coefficient	df	SS	MS	VC	% Variance	Coefficient	df	SS	MS	VC	% Variance
			GroupA						**GroupC**						**GroupE**			
P	-	22	47.13	2.14	0.0303	7.18	-	17	14.86	0.87	0.0192	4.34	-	27	12.738	0.472	0.0467	17.56
I	-	59	108.61	1.84	0.0659	15.61	-	35	156.64	4.47	0.2384	53.96	-	5	4.762	0.952	0.0272	10.22
PI	-	1298	578.70	0.33	0.3259	77.21	-	595	109.58	0.18	0.1842	41.69	-	135	25.905	0.192	0.1919	72.22
Phi	0.823						0.620						0.561					
G	0.848						0.789						0.593					
			GroupB						**GroupD**						**GroupF**			
P	-	23	30.79	1.34	0.0167	3.62	-	27	23.36	0.87	0.0178	4.57	-	27	13.542	0.502	0.0539	21.12
I	-	59	171.40	2.91	0.1070	23.17	-	35	150.96	4.31	0.1460	37.46	-	5	4.101	0.820	0.0229	8.99
PI	-	1357	458.74	0.34	0.3381	73.21	-	945	213.46	0.23	0.2259	57.97	-	135	24.065	0.178	0.1783	69.89
Phi	0.692						0.739						0.616					
G	0.748						0.632						0.645					

P, person; I, item; Phi, dependability coefficient; G, generalizability coefficient; df, degrees of freedom; SS, sum of squares; MS, mean square; VC, variance component.

Supplementary data 1. Generalizability study of non-PPI domains in 2023 and 2024, excluding reliability-hindering stations

2023							2024					
Effect	Coefficient	df	SS	MS	VC	% Variance	Coefficient	df	SS	MS	VC	% Variance
				Group A[†]						**Group C**[‡§]		
P	-	22	17.06	0.78	0.003	1.69	-	17	12.50	0.74	0.0026	0.53
C	-	9	50.22	5.58	0.012	6.74	-	7	22.63	3.23	0.0037	0.76
I:C	-	154	138.73	0.90	0.034	19.10	-	180	287.69	1.60	0.0799	16.40
PC	-	198	43.99	0.22	0.006	3.37	-	119	91.01	28.59	0.2402	49.31
PI:C	-	3388	418.25	0.12	0.123	69.10	-	3060	491.58	0.16	0.1607	32.99
Phi	0.501						0.541					
G	0.682						0.667					
				Group B						**Group D**[§]		
P	-	23	38.37	1.67	0.008	3.92	-	27	14.81	0.55	0.0019	1.12
C	-	11	61.93	5.63	0.012	5.88	-	7	15.67	2.24	0.0015	0.89
I:C	-	171	197.78	1.16	0.043	21.08	-	172	210.04	1.22	0.0393	23.27
PC	-	253	57.40	0.23	0.006	2.94	-	189	39.96	0.21	0.0040	2.37
PI:C	-	3933	531.81	0.14	0.135	66.18	-	464	566.95	0.12	0.1222	72.35
Phi	0.731						0.538					
G	0.850						0.610					

[†] One OSCE station (a combination of the pulmonary function test [PFT], meridian point pressure test [MPP], and Mini-Mental State Examination [MMSE]) was excluded because it yielded exceptionally low scores, possibly owing to difficulties that could not be resolved within the allotted time.

[‡] One OSCE station (a combination of acupuncture, chuna, and splinting) was excluded because it yielded exceptionally high scores, possibly because these procedures were relatively simple.

[§] One CPX station (headache) was excluded because of an excessively large between-day discrepancy, possibly resulting from first-day examinees disclosing the case to second-day examinees.

References

1. Tavakol M, O'Brien D: **Psychometrics for physicians: everything a clinician needs to know about assessments in medical education**. *International journal of medical education* 2022, **13**:100.
2. Chae H, Lee M-j, Kim M-H, Kim K, Cho E: **Multifaceted validity analysis of clinical skills test in the educational field setting**. *The Journal of Korean Medicine* 2024, **45**(1):1-16.
3. Mun SH, Jeong A-r, Lee H-Y: **A Scoping Review on the Current Status of Comprehensive Skill Competency Assessment in Korean Medicine Education**. *The Journal of Churna Manual Medicine for Spine and Nerves* 2025, **20**(1):57-72.
4. Jeong A, Lee H-Y, Shin S, Im S: **Challenges and solutions in developing an objective and structured clinical examination for complementary and integrative medicine: A mixed-method approach**. *Medicine* 2024, **103**(31):e39058.
5. Khan KZ, Ramachandran S, Gaunt K, Pushkar P: **The objective structured clinical examination (OSCE): AMEE guide no. 81. Part I: an historical and theoretical perspective**. *Medical teacher* 2013, **35**(9):e1437-e1446.
6. Manciu M, Trevino R, Mulla ZD, Cortez C, Plavsic SK: **Detection of biased rating of medical students by standardized patients: opportunity for improvement**. *Medical science educator* 2017, **27**(3):497-502.
7. Peeters MJ, Cor MK, Petite SE, Schroeder MN: **Validation evidence using generalizability theory for an objective structured clinical examination**. *INNOVATIONS in pharmacy* 2021, **12**(1):10.24926/iip. v24912i24921. 22110.
8. Pell G, Fuller R, Homer M, Roberts T: **How to measure the quality of the OSCE: A review of metrics–AMEE guide no. 49**. *Medical teacher* 2010, **32**(10):802-811.
9. Dwivedi NR, Vijayashankar NP, Hansda M, Dubey AK, Nwachukwu F, Curran V, Jillwin J: **Comparing standard setting methods for objective structured clinical examinations in a Caribbean medical school**. *Journal of Medical Education and Curricular Development* 2020, **7**:2382120520981992.
10. Khan U: **Standard setting OSCE: A comparison of arbitrary and Hofstee methods in a low stake OSCE**. *The Asia Pacific Scholar* 2024, **9**(3):15.
11. Jo HJ, Min SH: **The current status and future operations of Clinical Performance Evaluation (CPX) in the nationwide colleges (graduate schools) of Traditional Korean Medicine**. *The Journal of Korean Medical History* 2020, **33**(2):9-21.
12. Jo HJ, Roh JD, Sung HK, Park JS: **A survey on students' perception of clinical performance examination (CPX) in college of Korean medicine using student standardized patients.** *Journal of Society of Preventive Korean Medicine* 2020, **24**(3):1-13.
13. Cho E, Han Y-M, Kang Y, Kim J-H, Shin M-S, Oh M, Cho N-G, Jung H-J, Leem J: **Implementation of objective structured clinical examination on diagnostic musculoskeletal ultrasonography training in undergraduate traditional Korean medicine education: an action research**. *Diagnostics* 2022, **12**(7):1707.
14. Jiang Z, Ouyang J, Li L, Han Y, Xu L, Liu R, Sun J: **Cost-effectiveness analysis in performance assessments: a case study of the objective structured clinical examination**. *Medical Education Online* 2022, **27**(1):2136559.
15. Hara S, Ohta K, Aono D, Tamai T, Kurachi M, Sugimori K, Mihara H, Ichimura H, Yamamoto Y, Nomura H: **Feasibility and reliability of the pandemic-adapted online-onsite hybrid graduation OSCE in Japan**. *Advances in Health Sciences Education* 2024, **29**(3):949-965.
16. Haviari S, de Tymowski C, Burnichon N, Lemogne C, Flamant M, Ruszniewski P, Bensaadi S, Mercier G, Hamaoui H: **Measuring and correcting staff variability in large-scale OSCEs**. *BMC Medical Education* 2024, **24**(1):817.
17. Andersen SAW, Nayahangan LJ, Park YS, Konge L: **Use of generalizability theory for exploring reliability of and sources of variance in assessment of technical skills: a systematic review and meta-analysis**. *Academic Medicine* 2021, **96**(11):1609-1619.
18. Lertsakulbunlue S, Kantiwong A: **Development and validation of immediate self-feedback very short answer questions for medical students: practical implementation of generalizability theory to estimate reliability in formative examination designs**. *BMC Medical Education* 2024, **24**(1):572.
19. Sturgis PW, Marchand L, Miller MD, Xu W, Castiglioni A: **Generalizability Theory and Its Application to Institutional Research. The AIR Professional File, Spring 2022. Article 156**. *Association for Institutional Research* 2022.
20. Wu J-W, Cheng H-M, Huang S-S, Liang J-F, Huang C-C, Yang L-Y, Shulruf B, Yang Y-Y, Chen C-H, Hou M-C: **Comparison of OSCE performance between 6-and 7-year medical school curricula in Taiwan**. *BMC medical education* 2022, **22**(1):15.

21. Tavakol M, Brennan RL: **Medical education assessment: a brief overview of concepts in generalizability theory**. *International Journal of Medical Education* 2013, **4**:221.
22. Monteiro S, Sullivan GM, Chan TM: **Generalizability theory made simple (r): an introductory primer to G-studies**. In., vol. 11: The Accreditation Council for Graduate Medical Education; 2019: 365-370.
23. Yeates P, Moult A, Cope N, McCray G, Fuller R, McKinley R: **Determining influence, interaction and causality of contrast and sequence effects in objective structured clinical exams**. *Medical Education* 2022, **56**(3):292-302.
24. Trejo-Mejía JA, Sánchez-Mendiola M, Méndez-Ramírez I, Martínez-González A: **Reliability analysis of the objective structured clinical examination using generalizability theory**. *Medical Education Online* 2016, **21**(1):31650.
25. Lertsakulbunlue S, Kantiwong A: **Assessing medication counseling skills during transitioning from pre-clinical to clinical years among medical students: using generalizability theory to optimize reliable pharmacology exam design**. *Pharmacy Practice* 2025, **23**(2):1-12.
26. Park SY, Lee S-H, Kim M-J, Ji K-H, Ryu JH: **Acceptability of the 8-case objective structured clinical examination of medical students in Korea using generalizability theory: a reliability study**. *Journal of Educational Evaluation for Health Professions* 2022, **19**.
27. Peng Q, Luo J, Wang C, Chen L, Tan S: **Impact of station number and duration time per station on the reliability of Objective Structured Clinical Examination (OSCE) scores: a systematic review and meta-analysis**. *BMC Medical Education* 2025, **25**(1):84.
28. Paul A, Leung D, Salas RME, Cruz TE, Abras C, Saylor D, Gugliucciello V, Nunn J, Gamaldo CE, Strowd RE: **Comparative effectiveness study of flipped classroom versus online-only instruction of clinical reasoning for medical students**. *Medical education online* 2023, **28**(1):2142358.
29. Park YS, Hyderi A, Heine N, May W, Nevins A, Lee M, Bordage G, Yudkowsky R: **Validity evidence and scoring guidelines for standardized patient encounters and patient notes from a multisite study of clinical performance examinations in seven medical schools**. *Academic Medicine* 2017, **92**(Supplement_1):S12-S20.
30. Daniels VJ, Bordage G, Gierl MJ, Yudkowsky R: **Effect of clinically discriminating, evidence-based checklist items on the reliability of scores from an Internal Medicine residency OSCE**. *Advances in Health Sciences Education* 2014, **19**(4):497-506.
31. Kwon C-Y, Lee S, Hwangbo M, Cho C, Shin S, Kim D-H, Jeong A, Lee H-Y: **Implementation strategy for introducing a clinical skills examination to the Korean Oriental Medicine Licensing Examination: a mixed-method modified Delphi study**. *Journal of educational evaluation for health professions* 2023, **20**.
32. Feigerlova E: **Applying generalized theory to optimize the quality of high-stakes objective structured clinical examinations for undergraduate medical students: experience from the French medical school**. *BMC Medical Education* 2025, **25**(1):643.
33. Yu J, Lee S, Kim M, Lee J: **Comparison of students' performance of objective structured clinical examination during clinical practice**. *Korean Journal of Medical Education* 2020, **32**(3):231.
34. Carraccio C, Englander R, Van Melle E, Ten Cate O, Lockyer J, Chan M-K, Frank JR, Snell LS: **Advancing competency-based medical education: a charter for clinician–educators**. *Academic Medicine* 2016, **91**(5):645-649.
35. Oh S, Kim WJ, Kim MY: **Correlations between Medical Students' Self Assessment of Communication Skills and Patient-Physician Interaction Assessed by Standardized Patients in Clinical Performance Examination**. *International Journal of Bio-Science and Bio-Technology* 2015, **7**(3):1-8.
36. Clauser BE, Balog K, Harik P, Mee J, Kahraman N: **A multivariate generalizability analysis of history-taking and physical examination scores from the USMLE step 2 clinical skills examination**. *Academic Medicine* 2009, **84**(10):S86-S89.
37. Chong L, Taylor S, Haywood M, Adelstein B-A, Shulruf B: **The sights and insights of examiners in objective structured clinical examinations**. *Journal of educational evaluation for health professions* 2017, **14**.
38. Blood AD, Park YS, Lukas RV, Brorson JR: **Neurology objective structured clinical examination reliability using generalizability theory**. *Neurology* 2015, **85**(18):1623-1629.
39. Lyons KM, Brock TP, Malone DT, Freihat L, White PJ: **Predictors of pharmacy student performance on written and clinical examinations in a flipped classroom curriculum**. *American Journal of Pharmaceutical Education* 2020, **84**(12):8038.
40. Laird-Fick HS, Chang C, Wang L, Parker C, Malinowski R, Emery M, Solomon DJ: **Assessing the growth in clinical skills using a progress clinical skills examination**. *PeerJ* 2020, **8**:e9091.

41. Sidhu NS, Fleming S: **Re-examining single-moment-in-time high-stakes examinations in specialist training: A critical narrative review**. *Medical Teacher* 2024, **46**(4):528-536.
42. Moonen-van Loon J, Overeem K, Donkers H, Van der Vleuten C, Driessen E: **Composite reliability of a workplace-based assessment toolbox for postgraduate medical education**. *Advances in Health Sciences Education* 2013, **18**(5):1087-1102.
43. Homer M: **Setting defensible minimum-stations-passed standards in OSCE-type assessments**. *Medical Teacher* 2023, **45**(10):1163-1169.
44. Roh H, Park KH, Park SY: **Verbal communication of students with high patient–physician interaction scores in a clinical performance examination assessed by standardized patients**. *Korean Journal of Medical Education* 2017, **29**(4):241.
45. Donnon T, Paolucci EO: **A generalizability study of the medical judgment vignettes interview to assess students' noncognitive attributes for medical school**. *BMC Medical Education* 2008, **8**(1):58.
46. Clauser BE, Harik P, Margolis MJ, Mee J, Swygert K, Rebbecchi T: **The generalizability of documentation scores from the USMLE Step 2 Clinical Skills examination**. *Academic medicine* 2008, **83**(10):S41-S44.
47. Lipnevich AA, Mattern K, Feddock C: **Formative assessment and feedback in medical education: A practical guide: AMEE Guide No. 189**. *Medical Teacher* 2025:1-20.
48. McIlroy JH, Hodges B, McNaughton N, Regehr G: **The effect of candidates' perceptions of the evaluation method on reliability of checklist and global rating scores in an objective structured clinical examination**. *Academic Medicine* 2002, **77**(7):725-728.
49. Khan U, Khan YN: **Correlation between task-based checklists and global rating scores in undergraduate objective structured clinical examinations in Saudi Arabia: a 1-year comparative study**. *Journal of Educational Evaluation for Health Professions* 2025, **22**.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- STROBEChecklistCrossSectional.docx