Content validity and inter-rater reliability of procedural skill checklists used in the online OSCE scoring management system

Ferika Indarwati¹, Yanuar Primanda², Fahni Haris³, Resti Yulianti Sutrisno²

ABSTRACT

Introduction: The use of online Objective Structured Clinical Examination (OSCE) scoring system in medical and nursing education is emerging. To ensure valid student's score, the OSCE check lists need to be assessed for its validity and reliability. This study aims to test the validity and reliability of several procedural check lists commonly used in nursing profession such as peripheral intravenous insertion, electrocardiogram placement, nasogastric tube insertion, urinary catheter insertion and oxygenation procedure.

Methods: The expert consensus was used to generate items, content validity index and inter-rater reliability was used to evaluate the validity and reliability of the check lists. Five experts assessed the content validity of the checklists and five raters used the check list to evaluate performance of 11 students. Data were collected from April to May 2022. The panel experts rated the content relevance of each instrument using a four-point rating scale. Item level and scale level content validity index were calculated. Inter-rater reliability was calculated using Fleiss Kappa.

Results: The item validity index for the five check lists reviewed showed relatively high content validity among experts. I-CVI for each tool was very good, ranging from 0.8 to 1. The average content agreement (S-CVI/Ave) and the universal agreement (S-CVI/UA) for each check list were also very good. The inter-rater reliability results indicated that the agreement among raters was ranged from moderate to very good/excellent. The lowest Kappa value was for the nasogastric insertion check list (0.40, 95% CI 0.40 – 0.41) and the highest Kappa value was for the oxygenation check list (1, 95% CI 0.99 – 1).

Conclusion: The face validity was reported as easy to understand and presented logically. Nonetheless, re-formating of some items and addition of details in the checklists are needed to avoid ambiguity, which could lead to confusion for the examiner and examinee.

Keywords: Content validity, Inter-rater reliability, Online, OSCE, Skill.


INTRODUCTION

OSCE is considered as an objective method of assessing medical students' clinical competence.¹ The OSCE system was first introduced by Harden¹ in 1975 and became extensively used for assessment of other health professionals' education including nursing. In nursing education, the OSCE is commonly used to assess nurses' procedural competencies such as peripheral intravenous catheter insertion, electrocardiogram placement, recording and interpretation, nasogastric tube insertion, oxygenation procedure and urinary catheter insertion. The traditional OSCE scoring process is commonly used paper-based scoring system. There are several drawbacks regarding the paper-based OSCE scoring system which results are not immediately available. This leads to a delay of feedback; putting the academic staff on pressure, as well as the low ability to moderate and audit examination results. Paper-based test also requires high production costs; which would put more pressure on the administration staff.² The results of a study conducted in Indonesia also supported the Jansiraninatarajan and Thomas's study.² Paper-based OSCE scoring system was costly, wasted a lot of papers and had several limitations such as it requires more time for preparation and calculating the scores. It was less efficient in terms of resources, time, and management.³

The OSCE evaluation procedure is affected by technological advancements. A number of research suggest the development of an electronic-based scoring system.⁴,⁵ At one of the higher education institutions in Ireland, an electronic OSCE program (online OSCE management information system) has been developed to handle OSCE examinations with a cohort of first-year undergraduate nursing students provided for two consecutive years.⁶ It was discovered that using electronic software made it possible to save and analyze both group and individual findings collectively, saving a significant amount of time. Systems that only permit the electronic submission of fully completed forms guard against possible data loss. Additionally, the feedback tool enables students to compare and enhance their performance by
providing immediate evaluations of their work. There was no difference in the mean scores or failure rates between participants in the TeleOSCE and live OSCE, according to a study done in the United States using Zoom platform to conduct a summative pediatric OSCE. Another research, carried out at a university in the United Arab Emirates, assessed the viability and efficiency of e-OSCE for college students studying family medicine, medicine, and surgery in the COVID 19 period. The study discovered that the e-OSCE was well-liked by both students and examiners. The authors do not clarify whether the checklists employed in the online system have been evaluated for validity and reliability, leaving holes that need to be filled, even though the study has shown that online OSCE is favored over traditional techniques.

Numerous published studies on the online OSCE system in Indonesia have been reported; but relatively little study has looked at the reliability and validity of the checklists utilized in the online OSCE system. Regarding the nature of procedural skills measurement in health professional education, a number of systematic reviews have been published. Numerous strategies for assessing procedural abilities in medical education were cited in the majority of evaluations. For instance, according to Jelovsek al in 2011, there were more than 30 instruments accessible in medical education, the majority of which were developed to assess procedural skills in surgical education. In general, the review identified that there is a lack of assessment validation for procedural skills in the nursing profession. However, a checklist is available to assess common procedures performed by nurses such as peripheral intravenous catheter insertion, electrocardiogram insertion, recording and interpretation, nasogastric tube placement, oxygenation procedures and urinary catheter placement; the validity and reliability of this particular checklist has not been tested in the context of the online OSCE assessment management system.

Decisions related to scoring are important for any assessment modality. It is imperative to ensure that the scores obtained from the OSCE reflect examinees’ true ability as closely as possible. Some of the check lists used to assess the skills (e.g peripheral intravenous catheter insertion) are dated and use traditional binary scoring systems that restrict raters’ options, leading to non-valid estimation of ability. The growing body of evidence related to certain procedures such as the peripheral intravenous catheter insertion indicates that an updated check list needs to be formulated to examine the skill that allows benchmarking to the current standard. In addition, the use of the online OSCE scoring systems reiterates the importance of validating and testing the check list reliability when it is used in the online platform. Standardized checklists have been shown to improve trainee performance for many common procedures. A valid check list is crucial to assess the quality of test items to ensure that they are performing as expected. These measures provide empiric validity evidence for the internal structure of scores obtained, helping to ensure that scores represent a fair and reliable estimate of the students’ ability. The development of ON-OSCE scoring system will not be complete without refinement of the check list used in the platform. Therefore, this study aims to validate several check lists such as peripheral intravenous insertion, electrocardiogram placement, nasogastric tube insertion, urinary catheter insertion and oxygenation procedure used in the ON-OSCE scoring system.

METHODS

Study Design

This study employed several approaches to validate the IV insertion check list e.g Delphi methods for content validity, face validity by Polit and Beck and inter-rater reliability.

Sample selection

Five experts from nursing school and hospitals were involved in the content and face validity study. Five raters and 11 students were involved in the inter-rater study. The inclusion criteria for participant included in each phase of this study is described as follows: Experts who work in nursing education institution (in a particular field) in Indonesia for more than ten years, have experiences as examiners in OSCE, Have experiences in using paper-based OSCE scoring systems. The inclusion criteria for the interrater reliability phase are: instructors/lectures who have experiences in using check lists in the Online OSCE examination as raters, Nursing student with any level of ability/year of study. The exclusion criteria are those who are unable to participate in the inter-rater study’s phase.

The Instruments/check lists evaluated

There were five procedural skill check lists being evaluated in the present study namely peripheral intravenous (PIVC) insertion, electrocardiogram (ECG) placement, nasogastric tube (NGT) insertion, urinary catheter (UC) insertion and oxygenation (O2) procedure. The check lists had been formulated by medical and surgical department in the School of Nursing UMY. However, in the past five years it has never been updated against the current guideline and checked for its content validity and reliability. The IV-catheter insertion checklist consisted of six sections namely pre-interaction (5 items), orientation (9 items), work/core procedure (31 items), termination (9 items), documentation (5 items) and soft skills (6 items). The electrocardiogram placement includes pre-interaction (5 items), orientation (8 items), work/core procedure (25 items), termination (9 items), documentation (3 items), soft skills (7 items). The NGT insertion check list consisted of pre-interaction (4 items), orientation (8 items), work/core procedure (22 items), termination (7 items), documentation (3 items) and soft skills (5 items). The urinary catheter procedure check list has six assessment components namely pre-interaction (8 items), orientation (9 items), work/core procedure (21 items), termination (7 items), documentation (5 items), soft skills (7 items). The oxygenation procedure check list consisted of pre-interaction (5 items), orientation (8 items), work/core procedure (17 items), termination (9 items), documentation (5 items) and soft skills (7 items). The check list has non-binary scoring system. The scoring is divided into three components: raw score (0=not done at all – 5=done perfectly), critical points (1=Not critical – 3=very
critical) and difficulties points (1=Not difficult – 3=very difficult) to facilitate more flexible scoring. The final score is calculated as follows: raw score x critical point x difficulty point.

**Data Assessment/Collection**

Data collection for content and face validity tests were conducted using online correspondences for example WhatsApp and email. Potential experts from particular field related to the check list being evaluated were invited to participate in this study. The lead authors made sure that a balance combination of expert from academic and clinical/hospitals were involved in the validation process. The panel expert should have a minimum of 10 years’ experience in the clinical nursing fields, familiar with the procedure being evaluated and have experience on using the online OSCE system. All potential content validator received information sheets and the content validator form. Completing and returning the form indicates agreement of participation in the study. Five experts assessed the content validity of each check list. Data were collected from April to May 2022. The content validity was conducted following an expert review process described by Polit and Beck. The panel was asked to rate each questionnaire on a four-point scale, where 1 = not relevant, 2 = somewhat relevant, 3 = moderately relevant, and 4 = very relevant. They were also asked to provide clarity recommendations for each question on a four-point scale, where 1 = unclear, 2 = revised item (major), 3 = revised item (minor), and 4 = kept item as is. The items/questions were rated as relevant (fairly relevant or very relevant) and recommended for consideration for use in the adapted instrument. Items rated as relevant (very, moderate, and somewhat relevant) and revised items (minor or major) were deemed to require modification. Items rated as irrelevant have been removed from the tool. The panel members also evaluated each questionnaire regarding feasibility, clarity, logical sequence, and formatting. Qualitative comments from the panel were also reviewed. Data collection for inter rater reliability was conducted in the Mini Hospital PSIK FKIK UMY as well as through the online platform e.g., Zoom or Ms Teams. Lectures/instructors who meet the eligibility criteria were invited to participate in this study. Information sheets and consent were sought prior the study commenced. Five observers rated 11 student’s performance in each skill. In total there were 55 students’ performance that were evaluated by raters. The ON OSCE scoring platform will be used to collect the score from each student.

**Statistical Analysis**

Data analysis was conducted using Excel and SPSS statistics for windows version 28. The I-CVI was calculated by the number of item rated 3 or 4 divided by the number of experts, the S-CVI/ Ave was determined by the average of all I-CVI, and the S-CVI/UA was computed by the number of items rated relevant by all panel divided by the number of items. Items with I-CVI equivalent to > 0.78 or higher are considered good enough to be included in the final tool (Polit & Beck, 2011). The multi-rater Kappa coefficient was then computed using Fleiss Kappa in SPSS. The interpretation of the Kappa values as follows: 0.000 – 0.20 (Fair), 0.41 – 0.60 (Moderate), 0.61 – 0.80 (Good/Substantial), and 0.81 – 1.00 (Excellent/Very good) agreement among raters. Qualitative comments from the panel were also summarised.

**RESULTS**

Findings from the content validity test for each section of each procedure indicated that the I-CVI for each item was very good, ranging from 0.8 to 1. The average content agreement (S-CVI/Ave) and the universal agreement (S-CVI/UA) for each of the check list were also very good. The results of the content validity index of each procedure is presented in Table 1.

The result of the inter-rater reliability in table 2, indicated that the agreement among raters was ranged from moderate to very good/excellent. The lowest Kappa value was for the nasogastric insertion checklist and the highest Kappa value was for the oxygenation checklist.

The results from the clarity and relevancy of the check list indicated that the panel experts agree that the check list is relevant and clear to be used. However, there were several comments and recommendations from the panel members for modification of several items in the check list, so that it will be clearer and unambiguous. For example: washing hand also can be done by using six steps hand rub. In the patient identification the question should include patients date of birth or medical registration number. Panel also recommended that the checklist should also include identification of correct catheter type and size, infusion

**Table 1. Summary results of the content validity index each section of each instrument.**

<table>
<thead>
<tr>
<th>Skills</th>
<th>I-CVI</th>
<th>S-CVI/AVE</th>
<th>S-CVI/UA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peripheral intravenous catheter</td>
<td>0.8 - 1</td>
<td>0.98</td>
<td>0.92</td>
</tr>
<tr>
<td>Electrocardiogram placement &amp; recording</td>
<td>0.8 - 1</td>
<td>0.99</td>
<td>0.95</td>
</tr>
<tr>
<td>Nasogastric tube insertion</td>
<td>0.8 - 1</td>
<td>0.99</td>
<td>0.93</td>
</tr>
<tr>
<td>Urinary catheter insertion</td>
<td>0.8 - 1</td>
<td>0.99</td>
<td>0.95</td>
</tr>
<tr>
<td>Oxygenation procedure</td>
<td>0.8 - 1</td>
<td>0.98</td>
<td>0.92</td>
</tr>
</tbody>
</table>

**Table 2. Fleiss Kappa results of the inter-rater reliability for each check list.**

<table>
<thead>
<tr>
<th>Check list</th>
<th>Kappa value</th>
<th>95% Confidence Interval</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peripheral intravenous catheter</td>
<td>0.66</td>
<td>0.47 – 0.85</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Electrocardiogram placement &amp; recording</td>
<td>0.68</td>
<td>0.67 – 0.69</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Nasogastric tube insertion</td>
<td>0.40</td>
<td>0.40 – 0.41</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Urinary catheter insertion</td>
<td>0.51</td>
<td>0.50 – 0.52</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Oxygenation procedure</td>
<td>1</td>
<td>0.99 – 1.00</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

Sample data contains 11 effective subjects and 5 raters for each check list.
sets and appropriate application of pharmacological or non-pharmacological pain reliever. Documentation should focus on the patient response during and after the peripheral intravenous catheter insertion.

In the urinary catheter checklist, panel recommended putting details on the size of the syringe used to inject sterile water to lock the catheter. Furthermore, panel also suggested adding details related to any obstructions/distention felt by the inserter during catheter insertion in the urethra. If distention detected, then inserter may consider using condom catheter. For the nasogastric check list, panel experts recommended putting detail instructions such as during the nasogastric tube insertion if the patient has urgency to vomit, the inserter should stop inserting the tube and give patient a chance to take a deep breath and recollect their composure before the inserter re-starting the insertion. For the ECG check list, panel suggested to add details about what type of information should be given by the examinee to patients in the termination phase (e.g. limited to the cardiovascular rhythm only). For further results, patients can obtain it from the physician. The panel also suggested to add automatic ECG recording procedure side by side with the manual ECG procedure to accommodate if the automatic ECG is used during the OSCE. For the oxygenation check list, experts recommended to update detail related to the humidifier used, the modern humidifier is now already pre-filled with sterile water, so it can directly be attached to the flowmeter. Overall, panel members also recommended to use active sentence, consistent throughout the check list.

DISCUSSION

The findings demonstrate the validity and dependability of the five checklists examined in this study for evaluating students’ competencies. The majority of the instrument’s items displayed content validity and strong rater agreement. Every item on the checklist was determined to be pertinent to the most recent standard operating procedures related to each procedure.21-24 The various techniques employed to create the instrument can be utilized to explain this. Focus groups were employed to come up with the instrument’s components, and they included nursing lectures, medical professionals, and stakeholders who provided a wide range of crucial factors to consider when inserting IV catheters. Regarding the subjective experience of undergoing this operation, the clinician’s engagement is crucial. As a result, the variety of people involved and their opinions might have contributed to greater content validity.25

According to the reliability assessment, there was moderate to good agreement among the raters. On the majority of the items in each of the instruments, the assessors generally agreed to a very high degree, which is indicated by a high proportion of agreement and kappa values, which suggests satisfactory reliability. This conclusion is consistent with that of earlier research that looked at the interreliability of procedural OSCE checklists used in medical education.21,26,27 There are a number of reasons for the variation in rater agreement. Using a nonbinary scoring system, for instance. In an OSCE, the inclusion of non-binary scoring in the procedural checklists gives raters more freedom to assess the clinical competence of the examinees. Especially for diverse demographics or station types that call for more subjective assessment, this may actually lower rater variability. However, this element may also widen the range of the student's score. 14 However, it is believed that the non-binary scoring method is more precise in capturing the student's degree of proficiency.27

The intricate interactions between social, cognitive, and environmental elements that have an impact on raters are another aspect that may contribute to the heterogeneity in rater agreement. There is evidence that modifications related to these elements may have an impact on rater ratings, yet many of these characteristics are difficult to regulate.14 It is clear, however, that standardized checklists can result in improvements to psychometric traits.28 Standardization can boost trustworthiness in evaluation. The environment in which students are evaluated, such as hospitals and simulation facilities, where the interior is frequently unchanged, or other circumstances, may make it impossible for examiners to judge specific elements. Even if one of the items cannot be scored, the rest of the item can still be graded and discussed with the students in relation to the assessment.12

Furthermore, the raw score (non-binary) in the check list allows examiner to differentiate full or partial marks should be awarded to examinees who attempted (but did not satisfactorily complete) the items in the checklists.16 The addition of criticality and difficulty scoring in the check list allow for a systematic adjustment to the final score based on the criticality and difficulty of items that the examinee able to perform during the OSCE. By adding these scores, it is expected that if the examinee did not miss a critical items/skill central to the procedure, they can get higher score. This will also decrease leniency in the scoring system. A more lenient approach may result in lower failure rates and difficulties in differentiating competence versus non competence students.27 Therefore, by adding the raw, criticality and difficulty scores, it is expected that the check list can help to overcome those problems.

As a result of the overall higher scores translating into a substantially higher cut-score, it is therefore comprehensible that the stricter the scoring procedure paradoxically resulted in a higher failure rate. There can be little agreement among raters as a result of this higher cut score. The non-binary scoring system’s comparatively high ratings for pupils whose performance was assessed to be “borderline” can be used to explain why there was a greater failure rate. This is so that the learner can still obtain credit for attempting a procedure or a check-list item, even if they weren’t completed properly.16 This is possibly expected given that the raters may have set high standards for the student’s proficiency in relation to what they regarded adequate performance in terms of procedural skills. Additionally, some things on the checklist could be more crucial than others, such as preserving sterility, which could endanger the patient. If pupils don’t follow this important step, the tougher grading system will fail them. The impact on pass-fail judgments is significant, indicating that when choosing scoring techniques, the pass-fail outcome must be carefully examined. The scoring methods may not have affected the total

scores other than reflecting changes in the scale. As a result, including a criticality score in the checklist may help to lower variance and failure rates.

Other important measures for an OSCE include discrimination. In this study, reliability test indicated that the checklist showed relatively high discrimination property. It can differentiate the student’s level of ability and classified them as fail or successful in performing the procedure, and the agreement of discrimination among raters is excellent. One hypothesis for this finding may be that the assignment of the non-binary scoring system for the raw, criticality and difficulty of the items. Thus, prediction of the true value of the student's competent level can be achieved. For instance, the scoring system was able to distinguish between procedural items that required a failed “attempt” and might have a negative impact on patient care (such as incorrectly interpreting an ECG or using non-sterile gloves during a sterile procedure) and those that were important but not absolutely necessary and did not endanger patients’ safety. As long as they are on the proper road, applicants might be able to make up for little differences in communication abilities or structured orals.

In nursing education today, healthcare professionals, authorities and politicians pay increasing attention to quality of care and patient safety. Therefore, nurses should be competent to do routine or common procedures such as intravenous and urinary catheter insertions to improve patient outcomes. To prepare competent nurses, education institution should have reliable tools to evaluate students’ competency levels, the instruments used must be proved to have satisfactory stability and provide accurate observer ratings. The checklist validated in the present study include procedures that are very common in nursing profession such as peripheral intravenous insertion, electrocardiogram placement, nasogastric tube insertion, urinary catheter insertion and oxygenation procedure. The checklists have undergone rigorous validation processes. To our knowledge, the present study is the first research evaluating the five checklists commonly used to assess nursing student procedural competency level. As such, the findings from this study provide reliable tools to assess students’ performance. Compared to the subjective assessments that are frequently made now, it can provide more trustworthy performance evaluation outcomes. Standardized testing throughout foundational education may enhance the standard of nursing care and patient safety. Small sample numbers employed in the inter-rater reliability test, however, restrict the current study. Each method was examined on just 11 students. The checklist’s psychometric qualities might be established and further evaluated by additional research employing a bigger sample size.

CONCLUSION

The five checklists that were tested in this study have shown a fair amount of psychometric validity and reliability. The checklists can be used in a consistent manner to evaluate nursing students’ proficiency in tasks including inserting peripheral venous catheters, placing electrocardiograms, inserting nasogastric tubes, inserting urinary catheters, and conducting oxygenation procedures. Hospitals and nursing education facilities may find the approved tools helpful in their supervision of students. Institutions of higher learning may confidently evaluate students’ performance; students can also utilize the tools for training; and over time, they might improve patient safety.

CONFLICT OF INTEREST

The authors affirm that they have no known financial or interpersonal conflicts that would have seemed to have an impact on the research presented in this study.

FUNDING

The researcher acknowledges Universitas Muhammadiyah Yogyakarta that provides funding for this study.

ETHIC APPROVAL

Ethical clearance for this study was obtained from the Institutional review board with reference number: 127/EC-KEPK FKIK UMY/IV/2022.

REFERENCES


