ORIGINAL RESEARCH

Development of a valid and reliable evaluation instrument for undergraduate nursing students during simulation

Janeth J. Stiller *, Kristine A. Nelson, Mindi Anderson, Mary J. Ashe, Sharon T. Johnson, Kamal Sandhu, Ellen Mangold, Susan Scheid, Judy LeFlore

College of Nursing and Health Innovation, University of Texas at Arlington, Arlington, Texas, United States

ABSTRACT

Background: Many nurse educators are using instruments to evaluate undergraduate nursing students' performance during simulation. Rigorous reliability and validity testing for such instruments is often not presented. Evaluation and testing of instruments used is needed to support objective measurement of student performance.
Methods: This paper describes the development and testing of three new scenario-specific checklist instruments used for evaluating undergraduate nursing students' performance on an Objective Structured Clinical Examination using simulation. The new instruments were compared with original weighted ones previously used. An informal Delphi method was used to enhance content validity; Cronbach's alpha was utilized for reliability testing.
Results: Although neither set of instruments had been used with trained raters, the new set of instruments performed better on reliability analysis.
Conclusions: Checklist instruments may be more reliable in the objective measurement of student performance.

Key Words: Simulation, Undergraduate instruction, Reliable, Evaluation instrument, Nursing education, Competency

1. INTRODUCTION

1.1 Background

Measuring nursing students' application of knowledge, clinical competence, and skills performance using reliable, standardized methods is essential to the evaluation process and for practice.[1] To accomplish this, nurse educators are conducting simulation-based evaluations,[2] including those to assess performance.[3–7] Some are labeled as Objective Structured Clinical Examinations/Evaluations (OSCEs)[7–9] and others as high-stakes tests or evaluations.[10, 11] One of the benefits of patient simulations is that these can be specifically designed for students to demonstrate learning at the application, synthesis, and evaluation levels, as well as provide for a "realistic" yet safe environment for psychomotor skills performance and evaluation of values, attitudes, and beliefs.[6]

1.2 Review of literature

There is a growing interest in the use of health care simulation for summative assessment in which a student's performance is graded and level of competency determined.[11] A review of the literature reveals that nurse educators and researchers are using a variety of instruments to evaluate an individual student's performance during simulation and measure learn-

ing outcomes attained. However, few instruments, including those used in high-stakes testing like OSCEs, have had rigorous reliability and validity testing prior to utilization.[5] Others report that often in the literature, there is a lack of statistical support for reliability and validity of instruments that is presented.[1]

In 2010, Kardong-Edgren et al.[6] provided a review of evaluation instruments used in nursing simulations. Eight of these primarily assessed the cognitive learning domain; three assessed the psychomotor domain; and seven, the affective domain. Interrater reliability was reported for only two of the cognitive evaluation instruments; one of the psychomotor instruments; and one of the affective domain instruments. Most instruments were not evaluated for validity or had published validity data; four instruments reported content validity as determined by an expert faculty panel. Higher reliability was a consistent finding in studies that reported training of their raters.[6]

Adamson et al. (2013)[7] presented updated information on simulation evaluation instruments seen in the literature. One of their focuses was the discussion of instruments that evaluate OSCEs, some in undergraduate (UG) nursing. Reliability and validity data of the four instruments varied.[7]

In addition to having a valid and reliable instrument, it is imperative that careful consideration be given to selecting an instrument that is suitable to the activity and the learner.[7] Further, authors have suggested that researchers look to evaluate outcomes at a higher level, and that instruments be re-evaluated when used with new learners.[7]

1.3 Historical background

This study took place at a large Baccalaureate of Science in Nursing (BSN) program in the Southwestern United States. In August 2013, a new Associate Dean for Simulation and Technology determined that standardizing the implementation and evaluation of clinical simulation experiences in the undergraduate courses was essential. The second-semester senior Capstone faculty team volunteered to implement a plan for standardization.

Capstone is the final undergraduate clinical course taken by senior nursing students during their last semester and just prior to graduation. The focus of the course is "... on the synthesis of knowledge acquired throughout the curriculum, and the enactment of the professional nurse role, in a concentrated practicum" (p.2).[12] Capstone students must complete 96 clinical hours, or eight 12-hour hospital shifts, providing safe, comprehensive nursing care for a patient population in collaboration with and under the direct supervision of a registered nurse preceptor. Before starting the Capstone hospital

experience, each student must work individually to "pass" an OSCE test in the simulated hospital in order to be considered "ready" to start the Capstone practicum.

A committee consisting of most of the Capstone course faculty members (all undergraduate faculty), simulation faculty facilitators, and two faculty members with programs of research using simulation, was formed in spring 2014. At that time, Capstone course faculty members were using three instruments based on three different, standardized simulation scenarios for OSCE testing. A review of the three instruments by the two research faculty revealed that reliability and interrater reliability had never been established. Content validity was thought to be present based on faculty experience and consensus.

Historically, during OSCE testing, all Capstone faculty members used one of the three instruments based on the corresponding patient scenario. Capstone and other nursing program faculty members had developed these original instruments several years ago, with minor revisions made annually. Each instrument was divided into four categories including: Patient Safety, Assessment, Intervention, and Medication Administration; and each category was assigned a total score except for Medication Administration, which was pass/fail. Based on direct observation of a single student's performance, a Capstone faculty evaluator selected a numeric value from a range of possible points, for each weighted item. Because training of Capstone faculty evaluators had not been done prior to utilization of the three instruments, subjectivity in scoring was a concern.

Objective evidence is imperative to determining the competency level of the student. The weighted instruments included three different categories that allowed for diverse values when scored by individual faculty evaluators. Therefore, the committee decided to create new checklist instruments that could be scored as "Done" or "Not done" (D/ND). The aim was to compare reliability of the two sets of instruments, the original and new. Tables 1 and 2 are examples of one of the original and new instruments (based on one of the three corresponding scenarios) respectively.

2. METHODS

A proposal for tool development was submitted to the Institutional Review Board (IRB) at the university. It was determined that IRB approval was not needed.

2.1 OSCE scenarios

Students completed one of three possible OSCE standardized simulation scenarios. All three scenarios had been utilized in the Capstone course during previous semesters (with dif-

ferent students). Each scenario involved an initial clinical presentation in the form of change of shift report given by the Capstone faculty member. The scenarios presented adult patients with various medical-surgical conditions. After the student was briefed, he or she was expected to provide comprehensive nursing care appropriate to the patient's situation.

There was consensus among Capstone faculty and simulation facilitators that the scenarios were equivalent in complexity. Each faculty member assigned their students to one of three scenarios based on faculty member preferences and logistical factors; and each student performed her/his Capstone simulation individually.

Table 1. Example of original, weighted instrument (specific criteria not included)

Evaluation Criteria	Point Value	Done	Not Done	Points Earned
Patient Safety (30 points)				
1. Identifies self *(10 points)*	10			
2. Identifies patient:				
a. Name	5			
b. Date of birth	5			
3. Hand hygiene	10			
4. *Incorrect information given (-5 points)*				
Assessment and Nursing Diagnosis: (50 points) Initial VS for patient scenario				
5. Vital signs	*20 points*			
a. BP *(10 points)*	10			
b. Pulse *(5 points)*	5			
c. Respiratory rate *(5 points)*	5			
6. Focused assessment:	*20 points*			
a. Examines signs and symptoms *(5 points)*	5			
b. Auscultates breath sounds *(5 points)*	5			
c. Assesses cardiac system: auscultates *(3 points)*	3			
d. Inspects *(2 points)*	2			
e. Assesses IV *(5 points)*	5			
f. *VS taken (if not done: -5 points)*				
7. Nursing Diagnosis:	*10 points*			
Correct nursing diagnosis made *(10 points)*	10			
Planning & Interventions (20 points)				
8. Intervention: Correct *(10 points)*	10			
9. Intervention: Correct *(10 points)*	10			
Total Score: Must be 85% or above to pass				
Other: Incomplete – no score				
Medication Administration: PASS FAIL				
If student receives a "Not Done" for any portion of this section, it will result in a grade of "Fail" for the scenario				

Medication Administration	Done	Not Done
1. Right patient		
2. Assess patient for allergies		
3. Right drug		
4. Right dose		
5. Right route		
6. Right time		
7. Cleans top of vial		
8. Cleans IV port		
9. Use correct injection port		
Use correct injection port		

© University of Texas at Arlington, College of Nursing and Health Innovation

Table 2. Example of new instrument (specific criteria not included)

Evaluation Criteria Assign 1 point for "Done", 0 points for "Not Done"	Done	Not Done	Points
Patient Safety 5 points			
1. Identifies self to patient *(at least once)*			
2. Identifies patient:			
a. Name			
b. Date of birth			
3. Performs appropriate hand hygiene			
4. Safety measures			
Assessment & Nursing Diagnosis 13 points			
Initial VS for patient scenario			
5. Vital signs:			
a. Takes BP correctly			
b. Takes pulse correctly			
c. Counts respiratory rate correctly			
6. Focused assessment			
a. Assessment:			
• Identifies abnormal findings			
• Performs auscultation			
• Identifies patient breath sounds			
• Identifies patient symptoms			
b. Cardiac Assessment:			
• Auscultates			
• Identifies signs and symptoms			
• Identifies findings			
c. Recognizes patient complaints			
d. IV assessment			
e. Student recognizes need for intervention			
Interventions and Evaluation/Re-assessment (Critical Thinking & Decision-Making) 10 points			
7. Follows provider orders for correct intervention			
Evaluation/Re-assessment after intervention			
8. Identifies abnormal findings			
9. Auscultates			
10. Identifies lung sounds			
11. Checks provider orders			
12. Implements correct interventions			
Evaluation/Re-assessment after intervention			
13. Identifies correct findings			
14. Auscultates			
15. Identifies changes			
16. Notes VS changes			
Medication Administration 10 points			
17. Identifies right patient			
18. Assesses patient for allergies			
19. Administers right drug			
20. Cleans top of vial			
21. Administers right dose			
22. Cleans IV port			
23. Uses correct injection port			
24. Right time			
25. Right rate			
26. Maintains sterility			
Student points (out of 38 total):			

The three scenarios included: (1) a patient with heart failure named "Dakota Hampton", (2) a patient with an infection named "Ricky Samuels" and (3) a patient with respiratory distress named "Hayden Toner". The student was expected to assess and analyze key data within the OSCE, and provide appropriate interventions. Students had been previously exposed to the knowledge and skills required to be successful in the scenario; however, the components were never incorporated as a single summative or formative evaluation experience.

2.2 Instrument development

During the development of the new instruments, all faculty members endeavored to include essential elements of each nursing action, in accordance with best practice clinical guidelines and existing standards of care. After several iterations, the new instruments were finalized, and sent to the committee members for review and feedback. An informal Delphi method was used to enhance content validity: two external subject matter experts (SMEs) reviewed the instruments and provided feedback. Adjustments to the tools were made based on external SME comments.

For the three new checklist instruments, "Done" had a point value of 1 (one); and "Not done" a point value of 0 (zero). The sum for each of the four categories was added and a total score was obtained. The total score was based on a 100-percentage point system (see Table 2). Though involved in tool development, Capstone faculty members did not receive training prior to use of the new instruments.

Reliability of the original weighted instruments was evaluated in conjunction with the new instruments for each of the three patient OSCE scenarios. However, no decisions regarding pass/fail in evaluating the students were made using the new instruments.

2.3 OSCE protocol

The OSCE testing took place in the simulated hospital on campus, which is equipped with high-fidelity mannequins and a video/audio capture system. The Capstone faculty member remained in the room with his or her student in order to brief the student and then evaluate the student's individual performance. Each student's OSCE was recorded to facilitate instrument validity and reliability testing.

After the solo student entered the OSCE room, the faculty member read a set of standardized instructions to that student. Guidelines for reading these instructions were provided to all faculty members to maintain consistency.

Each student was given 30 minutes to complete the selected scenario; one-hour time slots were allotted for the student's

performance/evaluation and debriefing. The faculty agreed to refrain from speaking during the OSCE until the student finished the scenario. Following the student's completion of the scenario, debriefing was done at bedside by the faculty regarding the student's scenario-specific performance.

The faculty filled out the original scenario-specific, weighted tool first as the student proceeded through the scenario. The score obtained using the original tool was entered as the student's official pass or fail grade for the OSCE. Immediately following the scenario, the faculty completed the new scenario-specific checklist tool. Three blinded reviewers (two faculty members with experience in simulation and simulator research, who were involved with the committee, and one Graduate Research Assistant [GRA]) watched a live stream or reviewed recorded videos and scored scenarios using the three new instruments.

3. RESULTS

Seven of the eight Capstone faculty members conducted student evaluations using both the original and new instruments; therefore, not all students in the course were evaluated with the original and new instruments. Each of the three blinded reviewers watched live stream or recorded videos of different scenarios and scored using the new instruments.

A total of 125 instruments, including both the original and new, were scored. The distribution of the scenarios scored using the original instruments (N = 49) by Capstone faculty is shown in Table 3. Table 4 shows the distribution of scenarios scored using the new instruments (N = 76) by Capstone faculty and the blinded reviewers.

Table 3. Number of OSCE original weighted instruments

Patient Scenario	N
Hayden Toner	23
Ricky Samuels	22
Dakota Hampton	4
Total	**49**

Table 4. Number of new checklist instruments

Patient Scenario	N
Hayden Toner	38
Ricky Samuels	30
Dakota Hampton	8
Total	**76**

Cronbach's alphas of the original and new instruments are included in Table 5. As shown, the Cronbach's alphas on two of the scenarios were higher with the new instruments. The sample was too small on the scenario of the patient Dakota Hampton to calculate a Cronbach's alpha.

Table 5. Reliability comparison of original and new instruments

Patient Scenario	Form (Original/New)	N	Cronbach's alpha
Hayden Toner	Original	23	.24
Hayden Toner	New	38	.62
Ricky Samuels	Original	22	.22
Ricky Samuels	New	30	.68
Dakota Hampton	Original	4	0
Dakota Hampton	New	8	0

Table 6 shows the score differences in each of the three patients by category and total score. As noted, for the patient Dakota Hampton, the categories of Assessment and Intervention were statistically significant. There was no statistically significant difference in pass/fail in Medication Administration (χ^2 = .6). No other statistically significant differences were noted.

Table 6. Distribution of scores by patient scenario and category

Patient Scenario	Form Original/New	N	X ± SD	P value
Hayden Toner				
Patient Safety	Original	23	29 ± 3.3	.08
	New	38	27 ± 4.6	
Assessment	Original		44 ± 8.0	.25
	New		42 ± 7.0	
Intervention	Original		16 ± 6.0	.06
	New		13 ± 5.0	
Total Score	Original		89 ± 6.1	.35
	New		89 ± 6.7	
Ricky Samuels				
Patient Safety	Original	22	27 ± 5.5	.13
	New	30	29 ± 2.3	
Assessment	Original		44 ± 5.0	.08
	New		41 ± 7.0	
Intervention	Original		15 ± 4.0	.44
	New		14 ± 6.0	
Total Score	Original		86 ± 7.1	.54
	New		84 ± 6.9	
Dakota Hampton				
Patient Safety	Original	4	30 ± 0.0	.10
	New	8	27 ± 3.3	
Assessment	Original		48 ± 1.5	.003*
	New		33 ± 7.1	
Intervention	Original		20 ± 0.0	.000*
	New		9 ± 3.0	
Total Score	Original		95 ± 6.9	.000*
	New		78 ± 8.1	

*Statistical Significance

4. DISCUSSION

An effort was made by faculty to establish valid and reliable instruments in order to accurately assess Capstone students' readiness to practice. Although the faculty-developed, original instruments had been in use for three years, reliability had never been established. Due to a lack of formal rater

training in the use of a weighted scoring system, there was a concern that subjectivity could enter into the evaluation process. According to Hayden et al.,[5] many of the instruments used to evaluate students and their performance in clinical are thought to be subjective in nature.

Rushforth[13] suggests that because of variations in scenarios, skills assessed, number and qualifications of evaluators, and methods of scoring, every new OSCE tool needs reliability and validity testing, even when using a tool with previously established reliability. Formal rater training, discussions, and agreement among raters regarding acceptable minimum behaviors and cut-off passing scores prior to the simulated experience are crucial in increasing the interrater reliability.[14, 15]

Reliability testing of the original and the new set of scenario-specific instruments indicated that the new instruments had better reliability despite no prior rater training (Cronbach's alpha of 0.62 vs. 0.24 for Hayden Toner scenario; and 0.68 vs. 0.22 for Ricky Samuels scenario; see Table 5); therefore, the question is raised as to whether or not checklist instruments may perform better, even without rater training. Large faculty teams, faculty turnover, and difficulty with scheduling opportunities for faculty training are challenges many schools of nursing face. An unweighted checklist tool may provide a more objective measure to evaluate student OSCE performance.

Limitations

There are several limitations to this study. First, there was no rater training done prior to the new tool being implemented/utilized. Although there was no difference between the instruments on mean total scores or mean scores for each of the categories, test/retest phenomena may have influenced the scoring of the new instrument, since it was scored immediately after the original instrument was scored. Also, an unrecognized desire of the faculty to score the student in a positive way may have entered into the scoring process. Another limitation was the small number of evaluations done for one of the OSCE scenarios. Although content validity was supported by expert opinion, no data were collected to validate the items in the tool. Additionally, only a subset of the Capstone faculty in one university participated.

5. CONCLUSION AND RECOMMENDATIONS

Reliability and validity of all instruments used in simulation should be obtained and presented.[6] In this paper, we reported the development and the evaluation of a new set of checklist instruments and compared them to an original set of weighted instruments. Although neither set of instruments had been used with trained raters, the new checklist instruments performed better on reliability analysis. Cronbach's alphas on two of the three patient scenarios were higher with the new instruments. These findings raise a question: Do faculty need less training when using a checklist instrument due to the objective nature of this type of instrument? More research is needed to evaluate faculty training needs related to the type of evaluation instrument used. As shown in Table 5, reliability could not be assessed on the Dakota Hampton scenario. Additionally, differences shown in Table 6 need to be interpreted with caution because of the small number of times the scenario was used.

Our next step is to modify the new checklist instruments based on observations of the raters, establish interrater reliability, provide training for all faculty evaluators, and re-analyze the performance of the new set of instruments. From our experience in developing simulated clinical scenarios for OSCEs, it is important that the scenarios are specific to the learning outcomes and take into consideration the student's level in the nursing program. Evaluation instruments must match the scenario and be specific enough to identify the essential performance requirements in order to decrease subjectivity of the rater. For instance, one might say "Performs appropriate hand hygiene on entry to patient room", and it is scored as "Done" or "Not done". However, if the expected performance is not clearly identified, such as use of gel, gel and gloves, soap and water, amount of time needed to wash, etc., subjectivity in scoring may occur.

This instrument development supports the recommendation of Adamson and Kardong-Edgren (2012)[3] that valid and reliable instruments for assessing students in simulation are needed to accurately assess student performance in the simulation environment and to build evidence for the use of simulation to facilitate learning. Prion and Adamson (2012)[16] suggest that the "reliance on untested instruments measuring indirect data provides little support for educators trying to make informed decisions about integrating simulation into a curriculum, and it prevents researchers from building a credible body of knowledge about simulation" (p. e193).

CONFLICTS OF INTEREST DISCLOSURE
The authors declare that there is no conflict of interest.

REFERENCES

[1] Cant R, McKenna L, Cooper S. Assessing preregistration nursing students' clinical competency: A systematic review of objective measures. International Journal of Nursing Practice. 2013; 19(2): 163-176. PMid:23577974 http://dx.doi.org/10.1111/ijn.12053

[2] Adamson KA, Gubrud P, Sideras S, *et al.* Assessing the reliability, validity, and use of the Lasater Clinical Judgment Rubric: Three approaches. Journal of Nursing Education. 2012; 51(2): 66-73. PMid:22132718 http://dx.doi.org/10.3928/01484834-20111130-03

[3] Adamson KA, Kardong-Edgren S. A method and resources for assessing the reliability of simulation evaluation instruments. Nursing Education Perspectives. 2012; 33(5): 334-339. http://dx.doi.org/10.5480/1536-5026-33.5.334

[4] Ashcraft AS, Opton L, Bridges RA, *et al.* Simulation evaluation using a modified Lasater Clinical Judgment Rubric. Nursing Education Perspectives. 2013; 34(2): 122-126. http://dx.doi.org/10.5480/1536-5026-34.2.122

[5] Hayden J, Keegan M, Kardong-Edgren S, *et al.* Reliability and validity testing of the Creighton Competency Evaluation Instrument for use in the NSCBN National Simulation Study. Nursing Education Perspectives. 2014; 35(4): 244-252. http://dx.doi.org/10.5480/13-1130.1

[6] Kardong-Edgren S, Adamson KA, Fitzgerald C. A review of currently published evaluation instruments for human patient simulation. Clinical Simulation in Nursing. 2010; 6(1): e25-e35. http://dx.doi.org/10.1016/j.ecns.2009.08.004

[7] Adamson KA, Kardong-Edgren S, Willhaus J. An updated review of published evaluation instruments. Clinical Simulation in Nursing. 2013; 9(9): e393-e400. http://dx.doi.org/10.1016/j.ecns.2012.09.004

[8] Cazzell M, Howe C. Using objective structured clinical evaluation for simulation evaluation: Checklist considerations for interrater reliability. Clinical Simulation in Nursing. 2012; 8(6): e219-e225. http://dx.doi.org/10.1016/j.ecns.2011.10.004

[9] McWilliam PL, Botwinski CA. Identifying strengths and weaknesses in the utilization of Objective Structured Clinical Examination (OSCE) in a nursing program. Nursing Education Perspectives. 2012; 33(1): 35-39. PMid:22416539 http://dx.doi.org/10.5480/1536-5026-33.1.35

[10] Bensfield LA, Olech MJ, Horsley TL. Simulation for high-stakes evaluation in nursing. Nurse Educator. 2012; 37(2): 71-74. PMid:22327531 http://dx.doi.org/10.1097/NNE.0b013e3182461b8c

[11] Willhaus J, Burleson G, Palaganas J, *et al.* Authoring simulations for high-stakes student evaluation. Clinical Simulation in Nursing. 2014; 10(4): e177-e182. http://dx.doi.org/10.1016/j.ecns.2013.11.006

[12] Norman S, Adegbola M, Aguilar D, *et al.* NURS 4350: Transition to Professional Nursing (Capstone) Spring 2014. [cited 2014 Jul 5]. Available from: http://www.uta.edu/nursing

[13] Rushforth HE. Objective structured clinical examination (OSCE): Review of literature and implications for nursing education. Nurse Education Today. 2007; 27(5): 481-490. PMid:17070622 http://dx.doi.org/10.1016/j.nedt.2006.08.009

[14] Gantt LT. Using the Clark Simulation Evaluation Rubric with associate degree and baccalaureate nursing students. Nursing Education Perspectives. 2010; 31(2): 101-105. Available from: http://www.nln.org/nlnjournal/ PMid:20455367

[15] Parsons ME, Hawkins KS, Hercinger M, *et al.* Improvement in scoring consistency for the Creighton Simulation Evaluation Instrument©. Clinical Simulation in Nursing. 2012; 8(6): e233-e238. http://dx.doi.org/10.1016/j.ecns.2012.02.008

[16] Prion S, Adamson K. Making sense of methods and measurement: The need for rigor in simulation research. Clinical Simulation in Nursing. 2012; 8(5): e193. http://dx.doi.org/10.1016/j.ecns.2012.02.005