The Effect of Evaluator Training on Inter- and Intrarater Reliability in High-Stakes Assessment in Simulation

Ann E. Holland, Jone Tiffany, Linda Blazovich, Deborah Bambini, and Vicki Schug

Abstract

AIM The aim of this study was to evaluate the effectiveness of a training intervention in achieving inter- and intrarater reliability among faculty raters conducting high-stakes assessment of clinical performance in simulation.

BACKGROUND High-stakes assessment of simulation performance is being adopted in nursing education. However, limited research exists to guide best practices in training raters, which is essential to ensure fair and defensible assessment.

METHOD A nationwide sample of 75 prelicensure RN program faculty participated in an experimental, randomized, controlled study.

RESULTS Participants completing a training intervention achieved higher inter- and intrarater reliability than control group participants when using a checklist evaluation tool. Mixed results were achieved by participants when completing a global competency assessment.

CONCLUSION The training intervention was effective in helping participants to achieve a shared mental model for use of a checklist, but more time may be necessary to achieve consistent global competence decisions.

KEY WORDS Clinical Competence – High-Stakes Assessment – Interrater Reliability – Nursing Education – Simulation

Nursing programs are expanding their use of clinical simulation as a means of summative and high-stakes assessment of student learning. Although formative assessment is performed to provide feedback to students, "summative simulated-based assessment is intended to determine students' competence in practice" (Oermann, Kardong-Edgren, & Rizzolo, 2016, p. 323). The definition of high-stakes simulation currently outlined in the International Association of Clinical Simulation and Learning (INACSL) Standards of Best Practice: Simulation^{SM} is "an evaluation process associated with a simulation activity that has a major academic, educational, or employment consequence" (INACSL Standards Committee, 2016, p. S41). High-stakes assessment is a form of summative assessment, but with greater consequences.

In 2014, 43 percent of simulationists and educators responding to a nationwide survey reported that they were currently using manikins or standardized patients for high-stakes assessment (Rutherford-Hemming, Kardong-Edgren, Gore, Ravert, & Rizzolo, 2014). Participants in a town hall meeting held at the 2014 INACSL Conference expressed that they viewed summative and high-stakes assessment as means to judge student clinical competence (Rutherford-Hemming et al., 2014). Assessment of clinical performance to determine minimal competence to practice is viewed as necessary to ensure new graduates are prepared for present day and future practice (Benner, Sutphen, Leonard, & Day, 2010; Kavanagh & Szweda, 2017).

The findings from the landmark National Council of State Boards of Nursing (NCSBN) study that investigated the use of simulation as clinical substitution emphasized the importance of using valid and reliable mechanisms to assess achievement and competence in practice (Hayden, Smiley, Alexander, Kardong-Edgren, & Jeffries, 2014). Complexity in evaluating competence was the impetus for a seminal project undertaken by the National League for Nursing (NLN) to evaluate the process and feasibility of using manikin-based high-fidelity simulation for high-stakes assessment in prelicensure RN programs (hereafter referred to as the "NLN high-stakes study"; Rizzolo, Kardong-Edgren, Oermann, & Jeffries, 2015). Findings from the NLN study demonstrated that training evaluators to agree on what competence looks like is critical in summative and high-stakes simulation assessments (Oermann et al., 2016).

Findings from the NLN study compelled a research team to design a training strategy for evaluators and to study its effect on achieving inter- and intrarater reliability. The study reported here focused specifically on: a) training raters in the use of an evaluation tool, b) developing a shared mental model (SMM) of competence as it relates to the specific expected performance behaviors of a simulation scenario, and c) improving inter- and intrarater reliability of evaluators in an effort to contribute to the evidence informing valid and reliable high-stakes assessment in simulation.

About the Authors *Ann E. Holland, PhD, RN, is a professor, Bethel University, St. Paul, Minnesota. Jone Tiffany, DNP, RN, CNE, CHSE, ANEF, is a professor, Bethel University. Linda Blazovich, DNP, RN, CNE, is an associate professor, St. Catherine University, St. Paul, Minnesota. Deborah Bambini, PhD, RN, WHNP-BC, CNE, CHSE, ANEF, is a professor, Grand Valley State University, Grand Rapids, Michigan. Vicki Schug, PhD, RN, CNE, is a professor, St. Catherine University.*

This study was supported by the National League for Nursing, with funding from Laerdal Medical. The authors acknowledge the work of their research assistant, Dorie Fritz, MSN, RN, and statistician, Joel Frederickson, PhD.

For more information, contact Dr. Holland at ann-holland@bethel.edu.

The authors declare no conflicts of interest.

doi: 10.1097/01.NEP.0000000000000619

BACKGROUND

The NLN has played a leading role in guiding educators in student assessment. The NLN Presidential Task Force on High Stakes Testing outlined guidelines for fair testing for nursing education (NLN, 2012). Central to these guidelines is the definition of *fair*: that "all test-takers are given comparable opportunities to demonstrate what they know and are able to do in the learning area being tested" (p. 3). The guidelines emphasize the responsibilities and ethical obligations of faculty to thoughtfully and collaboratively investigate and implement high-stakes assessment practices. In the context of assessment in simulation, the findings from the NLN high-stakes study reinforced the need for the guidelines (Rizzolo et al., 2015).

The NLN high-stakes study focused on end-of-program assessment of video-recorded student performances in clinical simulation using the Creighton Competency Evaluation Instrument© (CCEI). The study yielded low interrater reliability among 10 raters. The evaluation phase was subsequently revised to focus the assessment on one scenario and to better define expected performance behaviors for the CCEI competencies. The revisions resulted in better interrater reliability but still only in the fair and good ranges of Kappa and intraclass correlation coefficient (ICC) (Rizzolo et al., 2015).

The NLN high-stakes study utilized two methods of assessing student competence. The CCEI is a checklist evaluation tool that prompts raters to score discrete, observable actions or decision-making as performed or omitted. In addition, participant evaluators were asked to rate students as competent or not competent and explain their rationale for the decision. The competence decision is a type of global rating scale (GRS) in which "an expert provides a holistic rating of the overall performance" (Boulet & Swanson, 2004, p. 124).

Ilgen, Ma, Hatala, and Cook (2015) emphasize the relative flexibility of GRS compared to checklists, which require modification for separate skills. However, GRS may require more rater training than checklists. Ilgen et al. (2015) found the two methods had similar interrater reliability. According to Boulet and Swanson (2004), both types of scoring criteria can be used to obtain reasonably accurate assessments; in addition, despite criticisms of subjectivity with global scoring, experts with proper training can provide reliable and valid scores. An important question that emerged from the NLN high-stakes study was: "What are the best methods to train raters?" (Rizzolo et al., 2015, p. 302).

Eppich et al. (2015) noted that examples of rater training programs in health care are scarce. They describe a rater training protocol to assess team performance using simulation-based methodologies that included three training sessions over 16 weeks. The protocol consisted of a didactic component, a review of the basics of rating performance using the selected tool, and three rounds of rating practice followed by facilitated discussion of discrepancies in the rating scores. Interrater reliability increased from the second training session to the third training when evaluating the same pilot video segment; moreover, four weeks following the training, adjacent agreement (percentage of times that two raters agree to within one unit of the score) was 97 percent and 90.6 percent at the end of all independent rating of 42 simulation videos. Eppich et al. assert that, although "use of a tool shown to yield valid and reliable data is a critical first step… rater training to use the tool in a calibrated manner is equally important to achieve reliability" (p. 87).

High-stakes assessment of student performance in simulation may be open to bias, inconsistency, and lack of fairness without rigorous processes for rater training. Observing and evaluating performance can be prone to rating errors and biases, especially for complex skills or assessment settings (Feldman, Lazzara, Vanderbilt, & DiazGranados, 2012). The INACSL Standards of Best Practice: Simulation (INACSL Standards Committee, 2016) suggest that required elements for high-stakes assessment using simulation-based experiences include "trained, nonbiased objective raters or evaluators," "using a comprehensive tool," and having "more than one evaluator for each participant" (p. S27). However, there is variability in the definition of terminology, criteria, and levels of training in using evaluation tools (Kardong-Edgren, Oermann, Rizzolo, & Odom-Maryon, 2017; Oermann, Yarbrough, Saewert, Ard, & Charasika, 2009).

One strategy to facilitate consistent and fair assessment is the development of an SMM among faculty evaluators. An SMM is described as "individually held knowledge structures that help team members function collaboratively in their environments and are comprised of four attributes: content, similarity, accuracy, and dynamics" (McComb & Simpson, 2014, p. 1485). Eppich et al. (2015) noted that, through training sessions, "raters developed a shared understanding of…skills and behaviors…and overall performance" (p. 89). An SMM enables faculty to have a more consistent and standard approach for student assessment (Boulet et al., 2011; Kardong-Edgren et al., 2017), which should lead to more fair and equitable assessment of student performance. The NLN high-stakes study research team reported the challenges of developing an SMM and stressed the importance of utilizing faculty with "similar values and professional judgment who are willing and capable of basing their judgments on the set criteria" (Kardong-Edgren et al., 2017, p. 66). They state, "Our findings demonstrate how important this preparatory work is when embarking on legally defensible high-stakes testing" (Kardong-Edgren et al., 2017, p. 67). This article describes a study extending the work of the seminal NLN high-stakes study, which sought to build an SMM through evaluator training for simulation performance assessment.

METHOD

Study Design and Instruments

The study was initiated after approval from the institutional review boards of the three universities at which research team members were employed. A nationwide study was conducted following completion of a pilot study to test and refine the tools and methods. An experimental, randomized, controlled design was used to investigate the effect of a training intervention on inter- and intrarater reliability.

Ten student performance videos of a clinical simulation scenario were selected, with permission, from the videos used in the NLN high-stakes study. Four members of the research team reached consensus in scoring these student performance videos using the form of the CCEI used in the NLN high-stakes study. One video demonstrating good student performance was used to orient participants to the CCEI and assessment procedures. Segments of the video with added evaluative audio commentary by a research team member served as an expert model for use in the training intervention. Three videos in each of three categories (representing good, mediocre, and poor student performance) were also selected.

Recruitment/Participants

Participants were prelicensure nursing faculty members with experience in clinical teaching and simulation recruited from across the United States through a recruitment letter, emails, the INACSL listserv, and personal contacts. Of 102 faculty who responded, 75 completed the study. The study retained equal numbers in control (n = 37) and intervention (n = 38) groups. There were no statistically significant differences between the groups in demographic characteristics such as age, years of teaching, or years of teaching with simulation.

Participants were first sorted into two groups based on whether they did or did not have experience with high-stakes testing. Each group was then randomized to the intervention or control groups. This sorting and randomization process controlled for the effect of rater experience with high-stakes assessment and the effect of training.

Training for Intervention Group

The research team developed basic orientation and advanced evaluator training (AET) modules (see Figure 1) that incorporated most elements of the training methodology established by Adamson and Kardong-Edgren (2012) to evaluate interrater reliability of the CCEI for use in the NCSBN national simulation study (Hayden, Keegan, Kardong-Edgren, & Smiley, 2014). The orientation and training modules were delivered to participants through a learning management system.

During the basic orientation, all participants completed a demographic survey, read printed materials about the study design and the CCEI, and viewed a video-recorded orientation to use of the CCEI for student performance assessment in simulation. Participants concluded the orientation by completing a practice assessment of one student performance video using the CCEI. Control group participants then proceeded to the experimental procedure of the study, described below.

After completing the basic orientation, intervention group participants began the AET. They participated with two to six participants in a training webinar led by a research team member. Participant scoring of the orientation video was reviewed during the webinar; an expert model interpretation of CCEI rating criteria for the simulation scenario was proposed at this time. The facilitator guided the discussion, helping participants reach agreement on scoring criteria and interpretations of competence. In the two weeks following the training webinar, participants independently rated three training videos of good, mediocre, and poor student performance using the CCEI, after which they participated in a small group coaching webinar to discuss their scoring results and reach consensus on an SMM of competent performance for the simulation scenario.

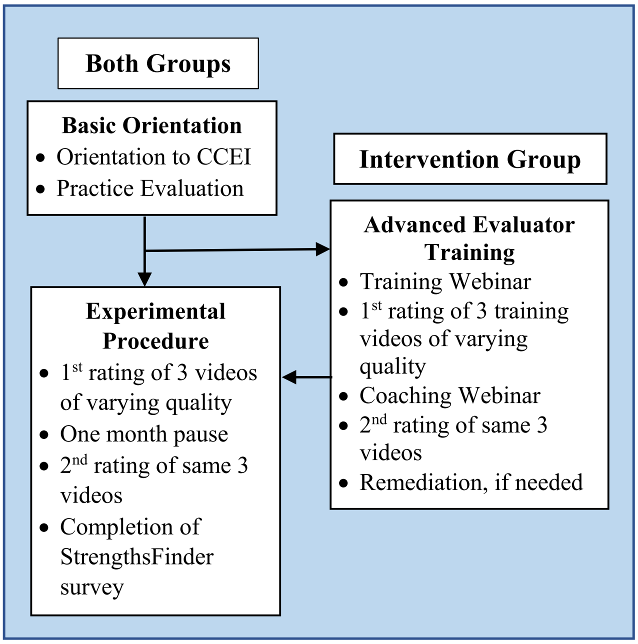
Both Groups
Basic Orientation
• Orientation to CCEI
• Practice Evaluation
Intervention Group
Advanced Evaluator Training
• Training Webinar
• 1st rating of 3 training videos of varying quality
• Coaching Webinar
• 2nd rating of same 3 videos
• Remediation, if needed
Experimental Procedure
• 1st rating of 3 videos of varying quality
• One month pause
• 2nd rating of same 3 videos
• Completion of StrengthsFinder survey

Figure 1. Study procedures.

Approximately one month after the first rating, the intervention group participants again rated the three training videos. Finally, individual telephone conference calls were conducted with intervention group participants whose second training video evaluations demonstrated patterns inconsistent with the SMM developed in the webinars. Individuals participating in this remediation call completed a final practice evaluation of a student performance video. Following the second evaluation of training videos and remediation if indicated, intervention group participants proceeded to the experimental procedure.

Experimental Procedure

Following the basic orientation (control group) or AET (intervention group), all participants completed the experimental procedure, which consisted of two ratings of three different videos (of good, mediocre, and poor performance) using the CCEI, separated by one month. Six videos were used in the experimental phase, with similar numbers of control and intervention group participants randomly assigned to rate each video. Ratings from the six videos were analyzed to determine inter- and intrarater reliability.

Data Analyses

Two scores reported by the participants were used as the dependent variables: the total score for the CCEI and the GRS competence decision. The ICC and Fleiss's Kappa statistic were used as measures of inter- and intrarater reliability. The ICC is routinely computed for continuous outcomes (CCEI scores). The data were analyzed using a multiple raters/measurements, consistency, two-way random effects model (McGraw & Wong, 1996). Kappa was also computed for the CCEI scores and as the single measure of inter- and intrarater reliability for the GRS competence decision (yes/no, binary measure). Kappa may better represent the effect of the training by correcting for chance agreement (Portney & Watkins, 2015). Kappa and ICC statistics were computed distinctly for the six experimental videos, rather than for the videos as a combined group, in order to examine if reliability was influenced by the quality of student performance.

RESULTS

Interpretation of ICC and Kappa values for inter- and intrarater reliability is accomplished through comparison to recommended scales. Accepted ranges for ICC include the following: <.40, poor reliability; .40 to .59, fair reliability; .60 to .74, good reliability; and .75 to 1.00, excellent reliability (Cicchetti, 1994). Portney and Watkins (2015) report accepted ranges for Kappa: <.40, poor to fair agreement; .40 to .60, moderate agreement; .60 to .80, substantial agreement; and >.80, high agreement.

CCEI Total Score

Table 1 summarizes interrater reliability statistics for the CCEI total score, comparing the intervention and control groups. The ICC statistics for the intervention group indicate excellent reliability (.869 to

Table 1: Interrater Reliability: Creighton Competency Evaluation Instrument Total Score

	Intervention Group		Control Group	
Video (NI, NC)	**ICC**	**Kappa**	**ICC**	**Kappa**
1 (19, 18)	.869	.769	.794	.527
4 (18, 20)	.969	.541	.738	.311
6 (20, 16)	.934	.411	.768	.294
11 (17, 19)	.922	.461	.771	.363
19 (21, 17)	.962	.658	.861	.584
28 (19, 18)	.952	.705	.567	.615

Note. NI = number of intervention group participants rating this video, NC = number of control group participants rating this video; ICC = intraclass correlation coefficient.

.969). Kappa statistics for the intervention group fall in the moderate to substantial ranges (.411 to .769). The ICC statistics for the control group are lower and range from fair to excellent reliability for the six videos (.567 to .861). Kappa statistics for the control group indicate poor to substantial reliability (.294 to .615). Overall, intervention group participants achieved higher interrater reliability, with less variability among the six videos.

Table 2 summarizes intrarater reliability statistics for the CCEI total score, comparing the intervention and control groups. The ICC statistics for the intervention group indicate excellent reliability (.811 to .957), whereas Kappa statistics fall in the substantial to high agreement categories (.723 to .951). Intrarater reliability for the control group is lower and more variable, with ICC statistics ranging from fair to excellent reliability (.576 to .953) and Kappa statistics ranging from moderate to high agreement (.503 to .829). Although both intervention group and control group participants achieved equally high intrarater reliability on some videos, the intervention group participants' intrarater reliability is more consistently high.

Competence Decision

Table 3 summarizes Kappa inter- and intrarater reliability statistics for the GRS competence decision. Intervention group interrater reliability indicates poor agreement (–.053 to .298); control group interrater reliability demonstrates variability between videos, ranging from poor to high agreement (–.009 to 1.00). Intrarater reliability is higher than interrater reliability for the competence decision for both groups. Kappa statistics for the intervention group indicate high agreement (.826 to .897). Kappa for the control group is again more variable, ranging from moderate to high agreement (.430 to 1.00).

DISCUSSION

The study results suggest that the structured training method, which included a model assessment, training webinars, practice assessments, and facilitated discussion, contributed to higher inter- and intrarater reliability in CCEI scoring. The intervention group achieved more agreement and less variability in scoring the CCEI than the control group. Computation of inter- and intrarater reliability statistics for each experimental video did not reveal a pattern indicating that reliability was influenced by the quality of student performance (good, mediocre, or poor).

This study used the same student video performances and CCEI tool as the NLN high-stakes study. The results of this study cannot be directly compared to the NLN high-stakes study because of different analytic techniques, but one can deduce that the training method used in this study produced higher inter- and intrarater reliability. The NLN high-stakes study reported fair interrater reliability (ICC = .58) when all 11 evaluators' scores and 28 videos were used and good interrater reliability (ICC = .62) when two problematic videos and two raters were removed from the analysis (Kardong-Edgren et al., 2017). Interrater ICCs for CCEI scoring in this study were predominantly in the excellent reliability range for both intervention and control groups.

The NLN high-stakes study reported good intrarater reliability when all 11 evaluators' scores and 28 videos were used (ICC = .70) and when two problematic videos and two raters were removed from the analysis (ICC = .73; Kardong-Edgren et al., 2017). This study produced intrarater ICCs in the excellent reliability range, with the exception of the control group for one of the videos.

An important study finding is that both intervention and control groups achieved high inter- and intrarater reliability for CCEI scoring. Even the control group ICC statistics were higher than the NLN high-stakes study, supporting the conclusion that the basic orientation was, in effect, a training intervention. The AET could be thought of as *intervention plus.* Components used in the basic orientation and AET have been found to be effective in evaluator training for clinical performance in studies conducted in nursing education and in other health professional education programs. For example, presentation of theory about assessment and the instrument were noted in two studies (Adamson & Kardong-Edgren, 2012; De Villiers & Archer, 2012). Practice ratings were used in nursing (Adamson & Kardong-Edgren, 2012), respiratory therapy (Rye, 2012), dentistry (Lin et al., 2013), and medicine (De Villiers & Archer, 2012; Lou et al., 2014), although some practice performances were live versus

Table 2: Intrarater Reliability: Creighton Competency Evaluation Instrument Total Score

	Intervention Group		Control Group	
Video (NI, NC)	**ICC**	**Kappa**	**ICC**	**Kappa**
1 (19, 18)	.811	.951	.912	.527
4 (18, 20)	.825	.885	.844	.541
6 (20, 16)	.858	.723	.770	.503
11 (17, 19)	.896	.821	.879	.522
19 (21, 17)	.957	.862	.953	.829
28 (19, 18)	.926	.891	.576	.616

Note. NI = number of intervention group participants rating this video; NC = number of control group participants rating this video; ICC = intraclass correlation coefficient.

Table 3: Inter- and Intrarater Reliability (Kappa): Competency Yes/No

	Interrater Reliability		Intrarater Reliability	
Video	**Intervention**	**Control**	**Intervention**	**Control**
1	.298	–.007	.826	.675
4	.268	1.00	.852	1.00
6	.211	.750	.875	.636
11	.029	.088	.876	.650
19	.010	.150	.897	1.00
28	–.053	–.009	.894	.430

video-recorded. Facilitated discussion following practice ratings was an effective component in some studies (De Villiers & Archer, 2012; Lin et al., 2013; Lou et al., 2014). Several studies incorporated feedback to raters after practice (Lin et al., 2013; Lou et al., 2014; Rye, 2012). In this study, all participants received presentation of theory about the CCEI, video-recorded instruction for applying the CCEI to the simulation scenario, and one practice rating. The elements unique to the AET contributed to higher inter- and intrarater reliability in the intervention group. For example, providing intervention group participants with a video-recorded model evaluation was a useful method to initiate an SMM in a group of evaluators from different regions of the country, practice specialties, and program expectations. Live discussions allowed participants to express their values and beliefs about clinical competence in attempts to reach consensus, whereas receiving feedback on their and others' ratings shaped their interpretation of scoring criteria. Participants, who were initially resistant, agreed through continued discussion and for some, remediation, to adopt an SMM.

Although results show the training was effective in shaping decisions about the CCEI scoring, the training was less effective in achieving inter- and intrarater reliability in the decision about students' overall competence. Inter- and intrarater reliability statistics were lower and more inconsistent for the competence decision than for the CCEI total score. In fact, the intervention group achieved consistently low interrater Kappas (–.053 to .298) and demonstrated less agreement than the control group (–.009 to 1.00) on most of the six videos. The NLN high-stakes study reported moderate interrater agreement (Kappa = .47) using 28 videos and 11 raters and substantial agreement (Kappa = .66) using only 26 videos and nine raters (Kardong-Edgren et al., 2017).

In contrast, the intervention group participants demonstrated high intrarater reliability Kappas for all videos (.826 to .897). Control group participants demonstrated higher agreement with themselves (intrarater) than with other participants (interrater), but control group intrarater agreement (.430 to 1.00) was highly inconsistent among videos. The NLN study reported substantial intrarater agreement (Kappa = .71) for the 26 videos and nine raters.

The competence decision findings prompted the research team to question if some cognitive dissonance developed within the intervention group participants as they were working toward an SMM. Did they adopt and integrate the SMM agreements for the discrete CCEI scoring, yet retain personal beliefs and values that conflicted with a shared agreement about a global competence definition? The study used a complex definition of competence: the ability to "observe and gather information, recognize deviations from expected patterns, prioritize data, make sense of data, maintain a professional response demeanor, provide clear communication, execute effective interventions, perform nursing skills correctly, evaluate nursing interventions, and self-reflect for performance improvement within a culture of safety" (Hayden, Jeffries, Kardong-Edgren, & Spector, 2009). During the AET webinars, significant discussion and disagreement occurred about what elements of the competence definition were most important. Some participants asserted that accurate clinical judgment was most essential (i.e., recognition of clinical changes, accurate identification of the problem, and appropriate intervention, including communication with the health care team). Other participants viewed elements of safety, such as handwashing and patient identification with two identifiers, as most essential. The training webinars may have provided inadequate time for participants to resolve these disagreements, recalling Ilgen et al.'s (2015) assertion that GRS may require more rater training than checklists. In addition, the prescribed nature of the model evaluation presented in the AET did not allow an organic development of a shared definition of competence that could more effectively develop within a team of raters who work together in a nursing program.

Limitations

Study completion required many hours of involvement in a structured timeline over a period of six weeks to three months, contributing to the attrition of one fourth of recruited participants. Many completing participants did not maintain the prescribed timeline, resulting in more than four weeks elapsing between the first and second video ratings. The greater elapsed time may have affected retention of knowledge gained during the orientation and AET, thereby influencing inter- and intrarater reliability.

This study used video performances created for the NLN high-stakes study. The accuracy of video performance assessment was highly dependent on the quality of audio and video capture and the structural limitations and variations of simulation labs in which the videos were recorded. Despite a careful evaluation process to select the best videos, some of the 10 videos selected had poor sound quality and visual capture. Some of the SMM agreements adopted by intervention group participants during the AET attempted to correct for structural and recording limitations of the simulation lab. For instance, in evaluating hand washing, it was agreed that the rater had to see the student wash/foam in, but the student would not be penalized if the recording stopped before they left the room and they did not foam out. Differences existed in cueing within the simulations. The videos varied in length; several were short enough to question whether the student had enough time to accomplish the expected tasks needed to score the CCEI. These weaknesses and variations of the video performances had the potential to confound the rating process of participants.

Implications for Research and Education

In the complex, quickly changing, and risky health care environment, assuring competence of practitioners is an obvious goal. It is clear that high-stakes assessment with simulation is coming to nursing education and practice. It is also clear that evaluator training is paramount to achieving fair assessment and that simulation that is utilized for high-stakes assessment must be designed, implemented, and facilitated according to standards of best practice. This study underscores some of the difficulty in achieving reliability in assessment and highlights the need for continued research.

Before the nursing profession can move forward with the implementation of high-stakes simulation assessment in a significant way, additional studies are needed to test and confirm the effects of training methods on reliability. The evidence-based training method used in this study could be replicated and further refined, perhaps with different evaluation tools. Continued study and refinement of evaluation tools is needed, recognizing that the objectives of the simulation must be consistent with the evaluation tool. The confounding results obtained in this study for the global competence decision indicate further study is needed in comparing and contrasting the use of checklists and GRSs in achieving reliable and valid assessments.

The experimental, randomized, controlled design and nationwide recruitment strengthen these study findings. Research that extends and replicates this study should use rigorous study designs and strive for larger participant samples. Such designs require funding, which is in short supply for nursing education research. Nurse researchers, administrators, and professional organizations such as the NLN must continue to advocate for change and forge partnerships with stakeholders in simulation and nursing practice to increase funding.

This study has significant implications for nursing education. Faculty and administrators who are using or considering the use of summative or high-stakes assessment in simulation must evaluate the resources, methods, and preparation invested in that assessment. Do program practices comply with evidence-based best practices, such as the INACSL Standards of Best Practice: Simulation, and emerging evidence about rater training? Are programs using psychometrically tested and validated assessment tools or untested home-grown tools? The results of this study lead us to concur with other researchers in urging caution in proceeding with high-stakes assessment at a program or nationwide level unless and until fairness and reliability is assured in our methods.

One of the big questions that must be answered in the movement toward high-stakes assessment is: How do nurse educators and professional organizations such as the NLN and the NCSBN build an SMM of clinical competence within programs and across the nation to ensure safe entry into practice? After evaluating the state of the science on clinical evaluation in nursing education, Lewallen and Van Horn (2019) called for nursing education to work toward standardized measures of competence, which can only be accomplished if competence can be consistently defined. Although this study utilized an accepted definition of competence, participants in this study judged student competence very differently, possibly because they prioritized components of the definition differently. Ongoing team efforts to achieve an SMM are needed among nursing faculty within a program to achieve a consistent and reliable judgment of competence. These discussions cannot be isolated from practice leaders since practice settings are where the “rubber meets the road.” Questions we would contribute to the discussion include: Is there reluctance in turning a low score on an objective tool into the rating “not competent”? Because nursing students are not fully formed practitioners at program completion, what elements of competence are nonnegotiable in end-of-program clinical performance? How does the global judgment of competence correlate with a checklist like the CCEI? How do we determine cutoff scores on checklist tools? Is the word *competent* so laden with meaning that we would be better served to use only the more objective scores? We invite readers of this article to engage in education/practice partnerships to seek answers to such questions.

CONCLUSION

This study demonstrated that a structured training method prepared faculty evaluators to achieve high inter- and intrarater reliability when using the CCEI. The training method was not as successful in helping participants to achieve an SMM about the global competence decision. The study provides nursing programs with a tested training method to guide faculty development. Researchers are urged to continue study in this area to inform methods and decisions to ensure student readiness for practice today and in future decades.

REFERENCES

Adamson, K. A., & Kardong-Edgren, S. (2012). A method and resources for assessing the reliability of simulation evaluation instruments. *Nursing Education Perspectives*, *33*(5), 334-339. Retrieved from www.nln.org

Benner, P., Sutphen, M., Leonard, V., & Day, L. (2010). *Educating nurses: A call for radical transformation*. San Francisco, CA: Jossey-Bass.

Boulet, J. R., Jeffries, P. R., Hatala, R. A., Korndorffer, J. R. Jr., Feinstein, D. M., & Roche, J. P. (2011). Research regarding methods of assessing learning outcomes. *Simulation in Healthcare*, *6*(Suppl), S48-S51. doi:10.1097/SIH.0b013e31822237d0

Boulet, J. R., & Swanson, D. B. (2004). Psychometric challenges of using simulations for high-stakes assessment. In W. F. Dunn (Ed.), *Simulators in critical care and beyond* (pp. 119-130). Des Plaines, IL: Society of Critical Care Medicine.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284-290.

De Villiers, A., & Archer, E. (2012). The development, implementation and evaluation of a short course in objective structured clinical examination (OSCE) skills. *South African Family Practice*, *54*(1), 50-54.

Eppich, W., Nannicelli, A. P., Seivert, N. P., Sohn, M.-W., Rozenfeld, R., Woods, D. M., & Holl, J. L. (2015). A rater training protocol to assess team performance. *Journal of Continuing Education in the Health Professions*, *35*(2), 83-90. doi:10.1002/chp.21270

Feldman, M., Lazzara, E. H., Vanderbilt, A. A., & DiazGranados, D. (2012). Rater training to support high-stakes simulation-based assessments. *Journal of Continuing Education in the Health Professions*, *32*(4), 279-286. doi:10.1002/chp.21156

Hayden, J. K., Jeffries, P. R., Kardong-Edgren, S., & Spector, N. (2009). *The national simulation study: Evaluating simulated clinical experiences in nursing education (Unpublished research protocol)*. Chicago, IL: National Council of State Boards of Nursing.

Hayden, J. K., Keegan, M., Kardong-Edgren, S., & Smiley, R. A. (2014). Reliability and validity testing of the Creighton Competency Evaluation Instrument for use in the NCSBN National Simulation Study. *Nursing Education Perspectives*, *35*(4), 244-252. doi:10.5480/13-1130.1

Hayden, J. K., Smiley, R. A., Alexander, M., Kardong-Edgren, S., & Jeffries, P. R. (2014). The NCSBN National Simulation Study: A longitudinal, randomized, controlled study replacing clinical hours with simulation in prelicensure nursing education. *Journal of Nursing Regulation*, *5*(2), S1-S66.

Ilgen, J. S., Ma, I. W. Y., Hatala, R., & Cook, D. A. (2015). A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. *Medical Education*, *49*(2), 161-173. doi:10.1111/medu.12621

INACSL Standards Committee (2016). INACSL standards of best practice: Simulation^{SM}. Participant evaluation. *Clinical Simulation in Nursing*, *12*(S), S26-S47. doi:10.1016/j.ecns.2016.09.009

Kardong-Edgren, S., Oermann, M. H., Rizzolo, M. A., & Odom-Maryon, T. (2017). Establishing inter- and intrarater reliability for high-stakes testing using simulation. *Nursing Education Perspectives*, *38*(2), 63-68. doi:10.1097/01.NEP.0000000000000114

Kavanagh, J. M., & Szweda, C. (2017). A crisis in competency: The strategic and ethical imperative to assessing new graduate nurses' clinical reasoning. *Nursing Education Perspectives*, *38*(2), 57-62. doi:10.1097/01.NEP.0000000000000112

Lewallen, L. P., & Van Horn, E. R. (2019). The state of the science on clinical evaluation in nursing education. *Nursing Education Perspectives*, *40*(1), 4-10. doi:10.1097/01.NEP.0000000000000376

Lin, C.-J., Chang, J. Z.-C., Hsu, T.-C., Liu, Y.-J., Yu, S.-H., Tsai, S. S.-L., … Lin, C.-P. (2013). Correlation of rater training and reliability in performance assessment: Experience in a school of dentistry. *Journal of Dental Sciences*, *8*, 256-260. doi:10.1016/j.jds.2013.01.002

Lou, X., Lee, R., Feins, R. H., Enter, D., Hicks, G. L. Jr., Verrier, E. D., & Fann, J. I. (2014). Training less-experienced faculty improves reliability of skills assessment in cardiac surgery. *Journal of Thoracic and Cardiovascular Surgery*, *148*(6), 2491-2496.e1-2. doi:10.1016/j.jtcvs.2014.09.017

McComb, S., & Simpson, V. (2014). The concept of shared mental models in healthcare collaboration. *Journal of Advanced Nursing*, *70*(7), 1479-1488. doi:10.1111/jan.12307

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30-46.

National League for Nursing. (2012). *Fair testing guidelines for nursing education*. Retrieved from www.nln.org/docs/default-source/default-document-library/fairtestingguidelines.pdf?sfvrsn=2

Oermann, M. H., Kardong-Edgren, S., & Rizzolo, M. A. (2016). Summative simulated-based assessment in nursing programs. *Journal of Nursing Education*, *55*(6), 323-328. doi:10.3928/01484834-20160516-04

Oermann, M. H., Yarbrough, S. S., Saewert, K. J., Ard, N., & Charasika, M. E. (2009). Clinical evaluation and grading practices in schools of nursing: National survey findings part II. *Nursing Education Perspectives*, *30*(6), 352-357.

Portney, L. G., & Watney, M. P. (2015). *Foundations of clinical research: Applications to practice*. Philadelphia, PA: F. A. Davis.

Rizzolo, M. A., Kardong-Edgren, S., Oermann, M. H., & Jeffries, P. R. (2015). The National League for Nursing project to explore the use of simulation for high-stakes assessment: Process, outcomes, and recommendations. *Nursing Education Perspectives*, *36*(5), 299-303. doi:10.5480/15-1639

Rutherford-Hemming, T., Kardong-Edgren, S., Gore, T., Ravert, P., & Rizzolo, M. A. (2014). High-stakes evaluation: Five years later. *Clinical Simulation in Nursing*, *10*(12), 605-610. doi:10.1016/j.ecns.2014.09.009

Rye, K. J. (2012). *Incorporating inter-rater reliability into your curriculum* [PowerPoint slides]. Retrieved from https://c.aarc.org/education/meetings/summer_forum/summer_forum.cfm

Instructions:

- Read the article. The test for this CE activity can only be taken online at www.NursingCenter.com.
  You will need to create (its free!) and login to your personal CE Planner account before taking online tests. Your planner will keep track of all your Lippincott Professional Development online CE activities for you.
- There is only one correct answer for each question.
  A passing score for this test is 13 correct answers.
  If you pass, you can print your certificate of earned contact hours and access the answer key. If you fail, you have the option of taking the test again at no additional cost.
- For questions, contact Lippincott Professional Development; 1-800-787-8985.

Registration Deadline: June 3, 2022.

Disclosure Statement:
The authors and planners have disclosed that they have no financial relationships related to this article.

Provider Accreditation:
Lippincott Professional Development will award 1.5 contact hours for this continuing nursing education activity.

Lippincott Professional Development is accredited as a provider of continuing nursing education by the American Nurses Credentialing Center's Commission on Accreditation.

This activity is also provider approved by the California Board of Registered Nursing, Provider Number CEP 11749 for 1.5 contact hours. Lippincott Professional Development is also an approved provider of continuing nursing education by the District of Columbia, Georgia, and Florida, CE Broker #50-1223.

Payment:

- The registration fee for this test is $17.95.