Journal of Nursing Measurement, Volume 25, Number 3, 2017

Testing Nursing Competence: Validity and Reliability of the Nursing Performance Profile

Janine E. Hinton, PhD, RN

Scottsdale Community College, Scottsdale, AZ

Mary Z. Mays, PhD

JMZ Enterprises, LLC, Tucson, AZ

Debra Hagler, PhD, RN, ACNS-BC, CNE, CHSE, ANEF, FAAN

Arizona State University, Phoenix, AZ

Pamela Randolph, BSN, MS, RN, FRE

PKR Nursing Education Consult, LLC, Phoenix, AZ

Ruth Brooks, MS, RN, BC, CHSE

Arizona State University, Phoenix, AZ

Nick DeFalco, MSN, RN

Scottsdale Community College, Scottsdale, AZ

Beatrice Kastenbaum, MSN, RN, CHSE

Arizona State University, Phoenix, AZ

Kathy Miller, MSN, RN

Scottsdale Community College, Scottsdale, AZ

Background and Purpose: There is growing evidence that simulation testing is appropriate for assessing nursing competence. We compiled evidence on the validity and reliability of the Nursing Performance Profile (NPP) method for assessing competence. Methods: Participants ($N = 67$) each completed 3 high-fidelity simulation tests; raters ($N = 31$) scored the videotaped tests using a 41-item competency rating instrument. Results: The test identified areas of practice breakdown and distinguished among subgroups differing in age, education, and simulation experience. Supervisor assessments were positively correlated, $r = .31$. Self-assessments were uncorrelated, $r = .07$. Interrater agreement ranged from 93% to 100%. Test–retest reliability ranged from $r = .57$ to .69. Conclusions: The NPP can be used to assess competence and make decisions supporting public safety.

Keywords: nursing competence; simulation testing; validity; reliability; practice breakdown

http://dx.doi.org/10.1891/1061-3749.25.3.431

Performance tests with purposes described clearly, validity supported by scientific evidence, and reliability demonstrated convincingly provide important information to support informed decision making about public safety, educational or employment access, and social justice initiatives (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). However, employers, regulatory officials, educational institutions, and certifying bodies who are responsible for assuring health care safety face challenges in creating tests to justly evaluate an individual health professional's clinical competence.

There are substantive barriers to objectively testing clinical competencies, including the responsibility to avoid patient risk when using authentic testing conditions (Meakim et al., 2013), the need to infer competence from a limited performance sample (Dreyfus & Dreyfus, 1980; Williams, Klamen, & McGaghie, 2003), and the requirement to assess a multidimensional construct involving characteristics that are not directly observable, such as cognition (Garside & Nhemachena, 2013). In addition, making valid inferences regarding competence based on samples of performance requires consideration of evidence from multiple independent sources (Williams et al., 2003). For the nursing workforce, the lack of a universally accepted definition of nursing competence adds another level of difficulty to evaluating clinical performance (Garside & Nhemachena, 2013).

High-fidelity simulation, already recognized as a useful teaching/learning method for health professionals, is increasingly being appreciated for its application in assessment and evaluation (Bensfield, Olech, & Horsley, 2012; Foronda, Liu, & Bauman, 2013; Whyte, Pickett-Hauber, Ward, Eccles, & Harris, 2013). Rizzolo, Kardong-Edgren, Oermann, and Jeffries (2015) demonstrated that carefully designed simulation scenarios employed in a controlled environment can support a psychometrically sound process to evaluate the clinical skills of prelicensure nursing students. However, limited research has been conducted using simulation for performance evaluation of nurses' continued competence.

The purpose of this article is to demonstrate that a rating instrument can be used in a valid and reliable manner to support the use of simulation in summative postlicensure testing. We summarize evidence collected on the validity and reliability of the Nursing Performance Profile (NPP) and show how the NPP can be used to quantify nursing competence, identify practice breakdown behaviors, and highlight demographic/professional characteristics of competency subgroups. We discuss methods for using the NPP to support decisions that protect public safety.

BACKGROUND AND CONCEPTUAL FRAMEWORK

Evaluation of clinical performance in authentic settings is possible using realistic simulations that do not place patients at risk. These assessments are more predictive of actual on-the-job competence than other assessment methods, such as written tests or performance records (Aebersold & Tschannen, 2013; Bensfield et al., 2012; Davis et al., 2006; Foronda et al., 2013; Randolph & Ridenour, 2015; Whyte et al., 2013). Considerable planning, piloting, and attention to detail are required to ensure that simulations used for formative or summative assessments present learners with opportunities to demonstrate the competencies being assessed and minimize factors that may bias performance (Meakim et al., 2013).

For more than 25 years, the Arizona State Board of Nursing (BON) sought valid, legally defensible measures of nursing competence to support decisions concerning nurses whose practice was being investigated (BON, 2007; Hinton et al., 2012; Randolph et al.,

2012). BON investigators have used multiple methods to develop a picture of the nurse's practice history including employer evaluations and counseling, respondent interviews, witness accounts, and patient records (Randolph & Ridenour, 2015). However, without an objective, direct measure of nursing competencies, the BON was challenged to justify its disciplinary decisions. Unable to identify competency measures that met all its objectives, the BON partnered with Scottsdale Community College and Arizona State University to develop a performance testing process using simulation that would provide a profile of nursing competencies. In the absence of a well-accepted instrument for assessing nursing competency during simulation tests (Aebersold & Tschannen, 2013; Kardong-Edgren, Adamson, & Fitzgerald, 2010; Whyte et al., 2013), we developed the NPP to allow a nurse to demonstrate competence in an easily observed, authentic, and safe setting. Competence was defined as the ability to meet minimum performance standards for safe practice in basic medical-surgical nursing care.

Recently, nursing competence was codified in the Quality and Safety Education for Nurses (QSEN) framework as a set of knowledge, skills, and attitudes organized into six categories (patient-centered care, teamwork and collaboration, evidence-based practice, quality improvement, safety, and informatics; QSEN, 2010). Furthermore, the Taxonomy of Error, Root Cause, Analysis and Practice-responsibility (TERCAP) system provided a structured approach to identifying patterns of error, risk factors, and system issues that contribute to practice breakdown (Benner et al., 2006). The NPP process uses these concepts of safe practice to identify instances of practice breakdown—any nursing practice action or lapse that is or might be harmful to a patient. Raters observe nurses and score their performance during simulated care using a high-fidelity manikin in a basic medical-surgical setting. The NPP rating instrument lists 41 tasks that comprehensively cover essential nursing competencies consistent with the QSEN and TERCAP systems.

PROCEDURES FOR INSTRUMENT DEVELOPMENT

Initial development of test content and format included reviews of the task domain, the congruence between the testing environment and the nursing workplace, and the relationships to other sources of performance data (Randolph et al., 2012). We created a rating instrument and three sets of three basic medical-surgical simulation scenarios. We conducted a Phase 1 pilot study (Hinton et al., 2012) to gather evidence on the feasibility and credibility of the testing scenarios; the reliability and validity of the rating instrument; and the process of having multiple experts independently score the videotaped performance of nurses completing simulation tests. The initial testing provided evidence on the basic feasibility, credibility, validity, and reliability of the NPP process. Raters were able to quantify the differences between nurses who engaged in consistent safe practice and those who did not, identify factors that contributed to high levels of competency, and highlight specific areas of practice that might respond to additional education and training (Hinton et al., 2012; Randolph & Ridenour, 2015).

Phase 2 of the project, the subject of this article, was a comprehensive assessment of the validity and reliability/precision of the final version of the test. We used the guidelines provided in the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014) to structure assessment and reporting of the validity and reliability/precision of the NPP process. The ways in which we addressed the enumerated standards of validity and reliability/precision are summarized in Tables 1 and 2, respectively.

TABLE 1. Summary of Evidence Provided on the Validity of the Nursing Performance Profile

Standard[a]	Interpretation[b]	Implementation[b]	Outcome[b]
1.1	The population of test takers and the constructs tested should be described, along with appropriate interpretation and uses of the test.	We defined the conceptual framework, test content, intended uses, simulation context, test takers, and raters.	The results and discussion delineated appropriate and inappropriate context, procedures, interpretations, uses, and consequences of testing.
1.2	The rationale for interpreting the test should be supported by theory and evidence.	We provided a brief review of literature on the need for, uses of, and barriers to high-stakes, summative testing using simulation. We explained the context within which we developed the test and how we envision it being used. We explained our approaches to providing evidence for its validity.	The results provided evidence congruent with the rationale for using simulation testing to assess nursing competence.
1.3	Potential inappropriate use of the tool should be defined.	We specified the conditions of use and assumptions underlying the test including the patient population, setting, and need for three independent evaluators.	The results provided evidence that the test is appropriate for measuring basic nursing competency when caring for medical-surgical patients.
1.4	New uses of the tool should be justified with new evidence.	Not applicable; the purpose of the current testing was to generate evidence for the original uses of the tool.	Not applicable.
1.5	The anticipated consequences of achieving a specific score should be supported by evidence.	We explained that the purpose of the test is to produce a profile of nursing competencies that testers can use to identify potential areas of practice for which a nurse might need additional education or training.	The Visual Nursing Performance Profile (NPP) illustrated a nurse's unique pattern of performance. When applied in a regulatory setting, it will provide a standardized view of a nurse's competence.
1.6	Explicit rationale should be provided to support anticipated indirect benefits.	Although we anticipate that testers can use the NPP to identify areas of practice breakdown, testing whether a nurse/employer/regulator could use the information to provide successful remediation was beyond the scope of the current project.	Not applicable

1.7	Practice effects should be evaluated.	We analyzed data from repeated testing to quantify the extent of practice effects in general and specifically for those without prior simulation experience.	Practice effects explained less than 1% of the variance in scores ($\eta_p^2 = .02$). *Note*: These results suggest that coaching for remediation will need to be more than simple familiarization with simulation testing.
1.8	Sample characteristics should be described.	We described our sampling methods for enrolling participants and raters. We compared the demographic and professional characteristics of the participants to state and national populations.	Participant demographics were congruent with state and national populations, with the exception that participants were more highly educated.
1.9	Raters' selection, training, and rating process should be described.	We described methods for selecting and training raters. We analyzed response patterns of raters.	Results confirmed that raters had more experience in nursing, simulation, and education. Response patterns confirmed that raters used the rating scale (*pass*, *fail*, *not observed*, *blank*) as instructed and were able to observe participants on all the tasks in the test domain.
1.10	Data collection procedures and settings should be described.	We described simulation testing scenarios, test settings, testing procedures, rating processes, and scoring in the current methods sections and by reference to prior publications.	Standardization of data collection procedures and settings enhances the ability to replicate investigations.
1.11	Procedures for generating test content should be described and justified.	We described the conceptual framework on which the test is based and the test development processes in the current methods sections and by reference to prior publications.	Test content was based on tasks required to provide safe care in an acute-care medical-surgical setting. Results confirmed that patterns of proficiency and deficiency matched those widely reported in the literature.

(Continued)

TABLE 1. Summary of Evidence Provided on the Validity of the Nursing Performance Profile *(Continued)*

Standard[a]	Interpretation[b]	Implementation[b]	Outcome[b]
1.12	When the rationale for scoring is based on premises about psychological/ cognitive processes of test takers or raters, theory or evidence to support it should be described.	We proposed that a valid measure of competence would be able to identify practice breakdown in areas known to be barriers to safe practice and further that the performance of test takers would vary with educational level and simulation experience.	Results confirmed that patterns of practice breakdown observed in the project were consistent with those widely reported in the literature and that they varied with educational level and simulation experience.
1.13	When the rationale for scoring is based on premises about the internal structure of the test, evidence on the internal structure should be described.	Not applicable; our exploratory analysis of the structure of the test is in progress.	Not applicable
1.14	When a scoring profile is recommended, evidence supporting the profile should be provided.	We explained the methods for deriving raw, profile, test, and overall scores and how they should be used.	Results confirmed that patterns of practice breakdown observed in the project were consistent with those widely reported in the literature and that they varied with educational level and simulation experience.
1.15	When interpreting scores on specific items is recommended, supporting evidence should be provided.	We explained the methods for deriving raw, profile, test, and overall scores and how they should be used.	Results confirmed that patterns of practice breakdown observed in the project were consistent with those widely reported in the literature and that they varied with educational level and simulation experience.
1.16	When a score/profile is analyzed for its relationship to a conceptually related variable, the rationale for selecting the variable should be provided.	We explained the rationale for comparing self-assessment ratings to rater scores.	Results indicated that self-assessment and rater scoring of videotaped performance were independent perspectives of competence, $r = .07$.

1.17	When a score/profile is analyzed for its relationship to a criterion variable, the rationale for selecting the criterion should be provided.	We explained the rationale for comparing supervisor assessment ratings to rater scores.	Results indicated that supervisor ratings and rater scoring of videotaped performance had a small, positive correlation, r = .31, in a subset of participants who were willing/able to obtain supervisor ratings.
1.18 and 1.19	When a score (alone or in conjunction with other variables) is expected to predict level of performance on a criterion, the relationship should be described.	Not applicable; we did not propose predictor–criterion relationships at specific levels.	Not applicable
1.20	When effect size measures are used to make inferences, they should be reported with indices of degree of uncertainty.	We used measures of effect size in conjunction with significance tests.	For example, the inferential test evaluating practice effects was not statistically significant and the effect size (partial eta squared) was near 0, $F(2,134) = 1.02$, $p = .36$, $\eta_p^2 = .02$.
1.21	Statistical adjustments should be explained.	We provided evidence of representative sampling and adequate dispersion of scores.	We did not need to make statistical adjustments.
1.22 and 1.23	Methods used in meta-analysis should be described.	Not applicable; we did not use meta-analysis.	Not applicable
1.24	When test scores are used to assign participants to different treatments, the evidence of differential outcomes should be reported.	Although we anticipate that testers can use the NPP to identify areas of practice breakdown, testing whether a nurse/employer/regulator could use the information to provide successful remediation was beyond the scope of the current project.	Not applicable
1.25	The root causes of unintended consequences of testing should be examined.	Not applicable; the purpose of the current testing was to generate evidence to support the intended uses of the tool. Evaluating unintended consequences was beyond the scope of the current project.	Not applicable

[a]Standards published in American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
[b]Information specific to the Nursing Performance Profile Project.

TABLE 2. Summary of Evidence Provided on the Reliability/Precision of the Nursing Performance Profile

Standard[a]	Interpretation[b]	Implementation[b]	Outcome[b]
2.1	Conditions for evaluating replications should be described and justified.	We explained the design of the project and the methods for data collection, repeated testing, standardization of testing procedures, and standardization of scoring (including simulation scenarios, testing environment, and rater training). We described approaches to evaluating raters, items, repeated tests, testing locations, and test sets.	Statistical results of comparisons across raters, items, tests, testing locations, and test sets indicated good standardization.
2.2	Methods chosen to evaluate reliability should be consistent with "replications associated with testing procedures" (p. 58[a]) and the intended uses of the test.	We evaluated different sources of measurement error from independent perspectives using different test scores and methods specific to the type of error.	Statistical methods and results provided quantitative estimates of measurement error associated with raters, items, repeated testing, testing locations, and test sets.
2.3	Indices of reliability/ precision relevant to each type of score should be reported.	We reported inter-rater reliability for raters' raw scores; internal consistency for Pass–Fail and Profile scores; test–retest correlations for Test Scores; and differences because of testing location and test sets for Overall and Profile scores.	Reliability/precision coefficients were consistently good. Inter-rater reliability was 93% or higher and was significantly higher than chance, $p < .00001$. Cronbach's α were .90 or higher. Pearson correlations for participants' repeated tests were .57 or higher, $p < .001$. Variability attributable to testing locations or tests sets was very small, $\eta_p^2 \leq .02$.
2.4	When interpretation of test scores is based on differences (gains/ losses over time or tests), standard errors should be reported.	Not applicable; we did not test the role of remediation interventions in this project. *Note*: We reported plots of performance across three tests (Visual Nursing Performance Profile [NPP]). However, we used the plots to illustrate raw scores for an individual or average raw scores for subgroups. The plots were not based on difference scores. Future projects could assess the *change* in the Visual NPP as a result of remediation interventions. Then, we would report standard errors.	

2.5	Reliability estimates should reflect the structure of the test.	Not applicable; our exploratory analysis of the structure of the test is in progress. In the interim, we matched methods for evaluating reliability to the type of score and the intended use of the score.	
2.6	Reliability coefficients should not be used interchangeably.	We used reliability coefficients relevant to each potential source of error and reported them separately.	See Outcomes for Standard 2.3.
2.7	When raters evaluate performance, evidence on both inter-rater reliability and participant performance across repeated measures should be included.	The description of methods explains how we trained raters and how they rated participant performance; specifically, three-person panels of raters scored repeated tests for individuals. We reported measures of both inter-rater and test–retest reliability.	Inter-rater reliability was 93% or higher. Pearson correlation coefficients for participants' repeated tests were .57 or higher.
2.8–2.18	Not applicable. These standards concern evaluating variability because of factors that we did not have (local scoring, short/long versions, local administration practices), were not relevant (age norms), were precluded by using repeated tests, or were incongruent with the intended use of the NPP.		
2.19	Methods should be carefully documented. Sampling procedures and descriptive statistics for samples should be reported. Reliability statistics should be described for each method.	We explained methods for sample selection, assessing sample characteristics, data collection, repeated testing, scoring, and assessing reliability/precision. We summarized sample demographics. We explained our approach to analysis of inter-rater reliability, internal consistency, test–retest reliability, reproducibility across testing locations, and reproducibility across testing sets.	Sample demographics were congruent with state and national populations. Statistical results indicated good reliability/precision. See Outcomes for Standard 2.3.
2.20	Statistical adjustments should be explained.	Not applicable; we did not need to make adjustments.	

[a]Standards published in American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
[b]Information specific to the Nursing Performance Profile Project.

DESCRIPTION, ADMINISTRATION, AND SCORING OF THE INSTRUMENT

Historically, validity and reliability were viewed as attributes of a test, which could be definitively established by a specific procedure or analytical technique. More recently, *validity* was defined as "the degree to which all of the accumulated evidence supports the intended interpretation of test scores for the proposed use" (American Educational Research Association et al., 2014, p. 26). Validity is based on the long-term collection of evidence from various settings about the test content, how it is administered and scored, and justification for how the scores are interpreted and used. Reliability/precision is based on evidence that demonstrates the "consistency of the scores across instances of the testing procedure" (American Educational Research Association et al., 2014, p. 47). Reliability/precision is a function of both the inherent variability of behavior and the degree of error in measuring behavior over time and across settings. Thus, background information about the content, administration, and scoring of the test is critical to understanding the evidence accumulated about the validity and reliability/precision of the NPP rating instrument. To facilitate the process of evaluating the quality of our evidence, we have linked our explanations to the standards codified by the American Educational Research Association et al. (2014).

To address Standards 1.1–1.3, 1.5, and 1.11, we refined our explanations of the test content, intended uses, context, and consequences of testing as we collected evidence. Our competency test was defined as a sample of a nurse's behavior in a simulated adult medical-surgical nursing setting that is scored using a standardized process. Our expectation was that the test results would be one source of information about a nurse's practice to be considered by decision makers who are charged with protecting the public's safety and to provide feedback to individual nurses about areas of practice in which they might benefit from additional education or training. It was our intention to provide a method to describe a nurse's performance and determine whether that performance was consistent with the minimum standard for safe practice. The test was not intended for use as a teaching method, to test the competency of groups of nurses, to estimate rates of practice breakdown in the population of registered nurses (RNs), nor to test specialty or advanced competencies of nurses. The test scenarios were developed for a simulated inpatient medical-surgical setting where care was provided to one manikin patient at a time. Results were not intended to predict competency in other health care settings or for multiple patient scenarios. To date, the test has been used with licensed and practicing RNs in Arizona; we have not evaluated its appropriateness for students or for nurses from other states or countries.

To address Standards 1.14, 1.15, 2.7, and 2.19, we standardized the rating, scoring, and score interpretation processes. A description of the 41-item instrument, and how raters used it, was tailed in Hinton et al. (2012). As in that previous study, staff in this study videotaped each participant completing a simulation test of medical-surgical nursing care in academic laboratories equipped for high-fidelity simulation testing. Raters scored each videotaped test on 41 competencies using one of four responses: *passed* (performance consistent with standards of practice), *failed* (exposed the client to risk for harm), *not observed*, or *left blank*. Four different scores were derived from the competency ratings: (a) the *Pass–Fail Score* on each competency item, (b) the *Profile Score* highlighting areas of proficiency and deficiency across the three simulation tests, (c) the *Test Score* summarizing performance across competencies on each test, and (d) the *Overall Score* summarizing simulation performance as a whole.

The Pass–Fail Score was computed from the raters' consensus about the participant's competence on a specific item. Nurses were rated as practicing unsafely only when two

or three raters rated the participant's performance on an item as *failed*; then, a consensus score of 0 = *failed* was assigned. Any other pattern of ratings was assigned a score of 1 = *passed*. That is, a passing score was assigned if two or three raters rated the competency item as *passed*, if two or three raters rated the competency item as *not observed*, or if the three raters disagreed on their ratings. In this way, a nurse was scored as failing only when multiple raters agreed there was clear evidence of unsafe practice.

The Profile Score was computed by adding the number of failures across the three tests into a final score of −3–0, where −3 meant the item was failed on all three tests, −2 meant the item was failed on two tests, −1 meant it was failed on one test, and 0 meant the item was never failed. These scores on each item were then plotted to provide a visual profile (Visual NPP) of an individual's nursing performance. An example Visual NPP is shown in Figure 1. Baseline 0 on the plot indicates minimally safe practice—the standard that all nurses should meet. Deviation below this standard indicates a practice breakdown. That is, the valleys on this plot should be interpreted in a commonsense way. When trained raters agree that a nurse has failed an item, and especially when this happens on more than one test, that item should be a target for additional training or education for the individual.

The Test Score was calculated for each participant on each testing scenario by computing the percentage passed (adding the number of items the participant passed, dividing the sum by 41, and multiplying by 100). Thus, a participant with a Test Score of 80% passed 33 of 41 competencies on that testing scenario.

The Overall Score is more stringent criterion of competency. It was calculated by computing the percentage passed on all three tests. Thus, a participant with an Overall Score of 90% passed the same 37 of 41 items in three different scenarios.

Test Scores and Overall Scores are useful in screening for individuals with low competency, because those low scores clearly indicate failures in multiple areas. When scores

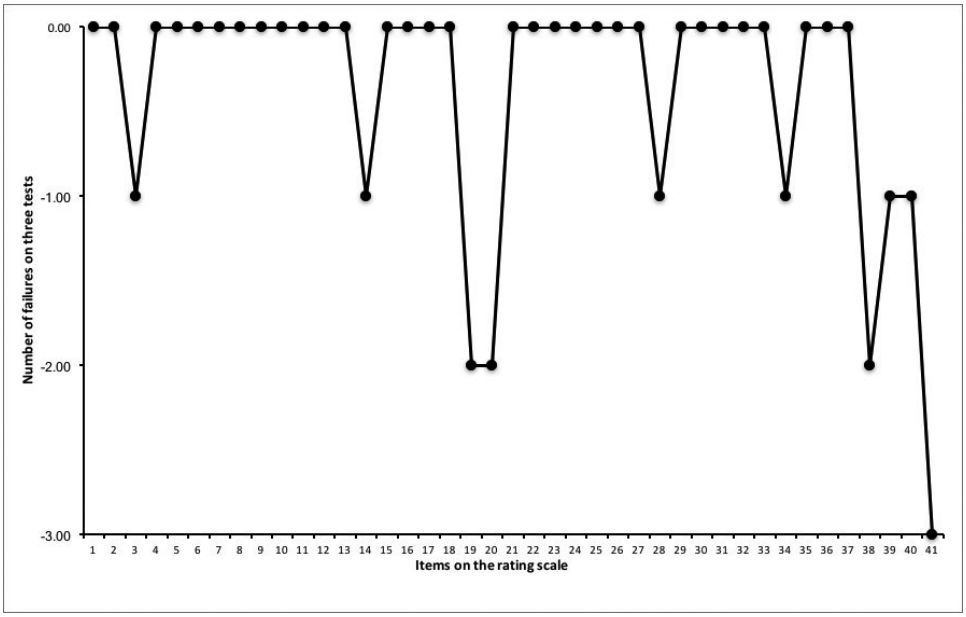

Figure 1. Example individual visual nursing performance profile.

are combined with the individual's Visual NPP and the raters' comments about reasons for their ratings, the tester has a detailed profile of the nurse's performance and the specific areas in which the nurse experienced a practice breakdown.

METHODS

Samples, Settings, and Procedures

We used sampling and data collection methods that would support inferences about the validity and reliability/precision of the NPP (Standards 1.8–1.10 and 2.1–2.3). The institutional review boards of both schools approved the project. All three organizations approved interagency and intergovernmental agreements. Participants ($N = 67$) and raters ($N = 31$) were recruited by e-mail invitations from the BON to all RNs licensed in the state, announcements in the BON newsletter, and an invitation on the BON website. Interested volunteers called a research phone line, were interviewed by phone, and scheduled for participation. Volunteers were recruited to serve as test participants and/or performance raters. For the raters, preference was given to those with at least a bachelor's degree in nursing, experience evaluating nursing practice, and three or more years of experience in nursing. Participants and raters provided written informed consent and were compensated for time and travel with $30 gift cards.

We used three sets of three common adult acute-care scenarios (Randolph et al., 2012) to test aspects of nursing care that all practicing RNs should be able to perform safely (NCSBN, 2007, 2008; QSEN, 2010). Testing scenarios included descriptions of the simulated patient's situation, background, and essential assessment data. Each scenario included a script to promote consistent verbal responses by the research staff portraying the patient, health care team, and visitors or family members. Preparatory information provided to participants 48–72 hr before testing included likely diagnoses, medications, treatments, laboratory tests, documentation forms, and instructions for programming the intravenous infusion pump. All tests were video recorded and archived for evaluation by raters.

We used simulation scenario sets in random order throughout the study. At the completion of the study, one set had been administered to 23 participants, generating 69 videos; the second set had been administered to 23 participants, generating 69 videos; and the third had been administered to 21 participants, generating 63 videos. Thus, 67 unique participants produced 201 videos for evaluation. Three raters, blind to the order of testing and the experience level of the participants, independently scored performance on each video (603 instruments; 24,723 item ratings).

Potential raters attended a 4-hr training session. The session included a summary of the NPP process, use of the NPP instrument, and practice in scoring performance videos. The raters reviewed a simulated patient's medical record before viewing the corresponding video. A copy of the state nursing practice act was provided as a resource for raters in assessing the performed scope of practice. During each training video, raters practiced using the NPP instrument to score the nurse's performance without discussion. After all trainees completed the NPP instrument for a particular video, the group discussed concepts of safe practice, rationales for assigned scores, interpretations of the scoring process, and performance items. Written comments were requested for any items scored as unsafe or not observed. Feedback from the raters was used to improve subsequent training sessions and videography. After the training, each rater scored up to 30 recorded simulation tests.

Approaches to Measuring Validity

The test was designed to assist the BON in making decisions about competence. Because nursing competency is based on education and experience (Benner, Sutphen, Leonard, & Day, 2010; Rosseter, 2015), our first approach to assessing validity addressed Standards 1.8–1.9 and 1.21 by evaluating whether the demographic and professional characteristics of our samples were congruent with those of the intended population of test takers and raters. We analyzed data using descriptive statistics.

One of the central threats to construct validity is a restriction of range in the test scores. We addressed Standards 1.8 and 1.21 using descriptive statistics to evaluate how well the test quantified the broad range of competence levels expected in a representative sample of volunteers.

Content validity requires that raters use the instrument as instructed and that the simulation test allow test takers to demonstrate all the competencies in the test domain. In this setting, that means raters must use all four possible ratings and use them in a nonrandom manner. To address Standards 1.9 and 1.11, we evaluated the response patterns of the raters using frequency analyses.

Tests that will be used to make high-stakes decisions must generate competence profiles consistent with familiar patterns of workplace performance and the existing literature on lapses in nursing practice with serious consequences for patient safety. We addressed Standards 1.11, 1.14, and 1.15 by compiling a frequency profile that highlighted the competencies on which most nurses were proficient and those on which most nurses were deficient and compared the profile to known strengths and weaknesses in nursing education and practice.

Safe practice in nursing is the minimum acceptable standard; it is based on well-established habits of assessment, critical thinking, and intervention within the scope of practice. Therefore, a competency test intended for working RNs should not show a practice effect with repeated testing. We addressed Standard 1.7 by comparing performance across the three simulation tests using one-way repeated measures analysis of variance (ANOVA). We used a mixed ANOVA to examine whether practice effects occurred in subgroups defined by previous simulation experience.

Although the literature on the validity and reliability of self-assessment methods is mixed (Dunning, Heath, & Suls, 2004; Ehrlinger, Johnson, Banner, Dunning, & Kruger, 2008), if self and observers' ratings are seen as two sources of evidence about competence, then evaluating their relationship should help to define the role of high-stakes competency testing. We used the NCSBN Clinical Competency Assessment of Newly Licensed Nurses (CCANLN) survey (2007) in comparison to NPP performance ratings. The CCANLN was used during Phase 1 of the project to develop the NPP testing instrument with permission to adapt survey items. The CCANLN consisted of 35 items that measured clinical competence, practice errors, and risks for practice breakdown through information reported by nurse–preceptor dyads using a scale (0 = *almost never*, 1 = *occasionally*, 2 = *fairly often*, 3 = *usually*, 4 = *almost always*, and NA = *no opportunity*). Per the CCANLN developers, reliability analysis revealed a Cronbach's alpha = .93 and content validity supported by content experts.

For this project, participants were asked to complete the CCANLN survey as a self-assessment of their nursing practice by rating themselves on the scale from 0 (*almost never*) to 4 (*almost always*) on each item. The sum of ratings could range from 0 to 140. A percentage *Self-Assessment Score* was calculated for each participant (by adding the

ratings across the 35 items, dividing the sum by the total possible score of 140, and multiplying by 100). This Self-Assessment Score indicated the participant's self-rating of his or her usual practice, with higher scores indicating higher levels of perceived competency. Computing the correlation between the Self-Assessment Score and the Overall Score addressed Standard 1.16. Supervisors in employment settings are expected to monitor and evaluate the performance of nurses; they are essential to protecting the public's safety. Despite controversy over the accuracy and precision of supervisor ratings (C. J. Gordon, Frotjold, & Bloomfield, 2015; Numminen, Leino-Kilpi, Isoaho, & Meretoja, 2015), a comparison of supervisors' ratings to the NPP should help to define the role of high-stakes competence testing. We asked supervisors to assess participants by rating the same CCANLN 35 items that were used by participants to complete self-assessments.

Each participant was asked to have a work supervisor complete the assessment of the participant's nursing practice. A percentage *Supervisor Assessment Score* was calculated for each participant. Computing the correlation between the Supervisor Assessment Score and the Overall Score addressed Standard 1.17.

In our project, some nurses consistently demonstrated the ability to meet the standard of zero errors across tasks on repeated tests; however, others failed to do so. We used the Overall Score to separate nurses into performance subgroups. Participants in the high-performance subgroup passed 36 or more items on all three tests ($n = 12$; top 20% of participants). Those in the moderate-performance subgroup passed 26–35 items on all three tests ($n = 42$). Those in the low-performance subgroup passed 25 or fewer items on all three tests ($n = 13$; bottom 20% of participants). We evaluated differences among subgroups using descriptive statistics and plots of the Visual NPP to address Standards 1.12 and 1.14.

Approaches to Measuring Reliability/Precision

We used a commonsense definition of inter-rater reliability based on raters observing a participant perform nursing care and agreeing with each other on whether the performance met safety standards. If ratings were random, then, over time, the four possible ratings (*failed*, *passed*, *not observed*, or *left blank*) would occur equally often. The rating process would be equivalent to each rater rolling a four-sided die. Independently, they would roll doubles and triples sometimes, but they would not do it 100% of the time. Conversely, if raters are using the ratings reliably and precisely, they should use some ratings more than others and agree nearly 100% the time, far more often than would be expected by chance alone. To evaluate inter-rater reliability, Standard 2.7, we calculated the proportion of instances in which two or three raters agreed on any one of the four ratings and used the binomial test to analyze whether, in a sample of 67 participants, the proportion was significantly higher than chance.

The 41-item rating instrument was designed to assess essential nursing competencies routinely required for safe nursing practice (Benner et al., 2006; NCSBN, 2007; QSEN, 2010). As such, the instrument should have a high degree of internal consistency. We addressed Standards 2.1–2.3 and 2.6 by calculating the Cronbach's alpha test for the Pass–Fail and the Profile scores.

The correlation among the three Test Scores is a measure of test–retest reliability. We addressed Standards 2.1–2.3, 2.6, and 2.7 by calculating the intercorrelation matrix of the bivariate Pearson correlation coefficients for Test Scores.

A reliable test should yield similar results regardless of where it is administered or which version of the test is administered. We addressed Standards 2.3 and 2.6 by comparing performance across (a) test locations using an independent samples *t* test on

the Overall Score and (b) three sets of tests using a one-way between-groups ANOVA on the Overall Score.

Approaches to Statistical Analysis

We analyzed data using SPSS (Version 22; IBM/SPSS Inc.). Congruent with Standards 1.20 and 2.19, we used descriptive and inferential statistics in combination with measures of effect size. We used descriptive statistics to illustrate patterns in the data. We used both parametric and nonparametric significance testing based on traditional Neyman–Pearson hypothesis testing to evaluate relationships among variables and performance differences among subgroups or across repeated tests. When multiple comparisons were made or the same test was repeated on multiple dependent variables, the Type I error rate (α) was adjusted using the Bonferroni correction of dividing the conventional $\alpha =$ by the number of comparisons. For example, computing a binomial test for inter-rater reliability on each item of the 41-item instrument required correcting the error rate for multiple comparisons (α is $.05/41 = .0012$). So each binomial test had to have a p value of .001 or less for the raters' agreement to be considered significantly different from chance levels.

Power analysis for the primary outcomes showed that a representative sample of 67 participants is sufficient to reject the null for meaningful effect sizes; for example, differences between subgroups of $d = 0.60$ with a power of 0.80, when α is set at .05 for a one-tailed t test. However, hypothesis testing is designed to disprove the null. Comprehensive validity assessment involves establishing not only that differences exist where they should but also that they do not exist where they should not. Thus, we report not only the achieved significance levels but also effect sizes. When effect sizes are extremely small (near 0), they are a good indication that no meaningful differences exist regardless of sample size.

RESULTS

Evidence Regarding Samples: Sample Characteristics

Evidence for validity and reliability/precision is more persuasive when the sample of participants represents the population of potential test takers and the sample of raters represents the desired pool of judges. The demographic and professional characteristics of participants and raters are summarized in Table 3. They were comparable to the population of RNs in Arizona and the United States. For example, recent state and national estimates of the percentage of nurses who were (a) 55 years old and older ranged from 28% to 53%, (b) men ranged from 7% to 11%, and (c) non-Hispanic White ranged from 70% to 90% (AMN Healthcare, 2013; Budden, Zhong, Moulton, & Cimiotti, 2013; Department for Professional Employees, 2015; Johnson et al., 2009; Randolph, 2016; Rosseter, 2014). The distribution of these characteristics in our participants fell into those ranges. However, the sample was more highly educated with 83% of participants having a bachelor of science in nursing (BSN) degree or higher compared to state and national estimates ranging from 45% to 63% (AMN Healthcare, 2013; Budden et al., 2013; Department for Professional Employees, 2015; Randolph, 2016; Rosseter, 2015). This demographic feature of the sample is in keeping with the goal of having 80% of the RN workforce have a baccalaureate degree by 2020 (Rosseter, 2015). By design, raters were more experienced in nursing, more highly educated, and more experienced with simulation than participants (see Table 3).

TABLE 3. Demographic and Professional Characteristics of the Sample

Attribute	Participants *n* (%)	Raters *n* (%)
Age		
Younger than 55 years	39 (58)	16 (52)
55+ years	27 (40)	14 (45)
Not answered	1 (1)	1 (3)
Gender		
Male	6 (9)	3 (10)
Female	61 (91)	28 (90)
Ethnicity		
Hispanic	3 (5)	0 (0)
Non-Hispanic	63 (94)	31 (100)
Not answered	1 (1)	0 (0)
Race		
White (non-Hispanic)	57 (85)	29 (94)
Black/African American	2 (3)	1 (3)
American Indian	1 (1)	0 (0)
Asian	1 (1)	0 (0)
Other/Not answered	6 (10)	1 (3)
Highest education		
Diploma	0 (0)	1 (3)
Associate	11 (16)	0 (0)
Bachelor	25 (37)	9 (29)
Masters	27 (40)	14 (45)
Doctorate	4 (6)	7 (23)
Simulation experience		
None	20 (30)	7 (23)
Occasional	30 (45)	13 (42)
Frequent	14 (21)	11 (36)
Not answered	3 (5)	0 (0)
Attribute	Participants *M* (*SD*)	Raters *M* (*SD*)
Age (years)	49.26 (11.88)	51.83 (8.56)
Years as RN	22.18 (12.08)	28.16 (10.20)

Note. RN = registered nurse.

TABLE 4. Dispersion in Summary Scores

Score	Min	Max	*M* (*SD*)
Test 1	20	100	84.24 (15.03)
Test 2	37	100	86.06 (12.61)
Test 3	34	100	86.02 (12.61)
Overall	12	98	72.55 (17.61)

Evidence Regarding Samples and Statistical Validity: Dispersion of Scores

Descriptive statistics on scores are shown in Table 4. It is clear that their dispersion adequately represented the possible range of performance expected among experienced, licensed RNs from a broad spectrum of nursing roles. The adequate dispersion of scores supports straightforward analysis and interpretation of scores without adjustment for restriction of range.

Evidence Based on Response Processes: Rater Response Patterns

Raters should use all four possible ratings (*passed*, *failed*, *not observed*, or *left blank*) and use some of them more often than others, congruent with the testing conditions and performance of the participants. Raters left the rating blank on an item up to 2% of the time or scored the item *not observed* up to 7% of the time. Across the 41 items, there was only 1 item that raters never *left blank* or rated *not observed* ("Performs within scope of practice"). There were three items that were rated *not observed* 6%–7% of the time. These items are among the most complex nursing competencies ("Recognizes when care demands have exceeded nurse's capacity," "Delegates/coordinates aspects of care appropriately," "Correctly records telephone orders"; Benner et al., 2010; Bensfield et al., 2012). These rater response patterns are critical evidence of rater integrity and content validity. The patterns indicate that raters were using all four ratings and that, with occasional exceptions, they were able to observe participants performing all the tasks in the test domain.

Evidence Based on Relations to Other Variables: Expected Patterns of Competency

Areas of nursing practice where nurses are routinely found to be competent are well described in the literature (Benner et al., 2010). Our results mirror these proficiency patterns. There were nine items that 90% or more of nurses passed on all three tests. Examples of these items are "Recognizes changes in client condition necessitating intervention," "Provides specific interventions tailored to client/family vulnerabilities," and "Provides respectful and culturally responsive care." These proficiency data provide evidence of construct validity. There were seven items that less than 50% of nurses passed on all three tests. Examples of these items are "Administers medications accurately and safely," "Demonstrates application of infection control principles," and "Documents accurate and legally defensible account of interventions." These deficiency data provide evidence of construct validity because the deficiencies occur on items that are among the most common lapses in safe practice (Benner et al., 2010; Bensfield et al., 2012). When these lapses occur, there are substantial threats to patient safety (Institute of Medicine [IOM], 2011).

Evidence for Intended Uses and Interpretations: Practice Effects

Our test is intended to be an examination of everyday performance, of fundamental competencies that are well learned, so there should be no practice effects during testing. A repeated-measures ANOVA on Test Scores 1, 2, and 3 did not show statistically significant improvement in scores with repeated testing (Table 5). That is, practice effects accounted for less than 1% of the variance in test scores ($\eta_p^2 = .02$ in Table 5). This evidence supports the interpretation of scores as a measure of current competency rather than ability to learn or adapt.

Although the sample as a whole did not show practice effects, participants without prior simulation experience might show a practice effect. A 3 × 3 mixed ANOVA comparing Simulation Experience Subgroups across Repeated Tests showed no significant main effect of repeated tests, nor was the interaction between simulation experience and repeated tests statistically significant (see Table 5). The very small effect sizes for these

TABLE 5. Differences Because of Practice Effects

One-Way Repeated Measures Analysis of Variance	
	Repeated Test *M* (*SD*)[a]
Test 1	84.24 (15.03)
Test 2	86.06 (12.61)
Test 3	86.02 (12.61)

Two-Way Mixed Analysis of Variance			
Simulation Experience			
	Subgroup *M* (*SD*)[b]		Repeated Test *M* (*SD*)[c]
Frequent ($n = 14$)	92.10 (5.46)	Test 1	90.59 (6.26)
		Test 2	93.38 (4.53)
		Test 3	92.33 (5.48)
Occasional ($n = 30$)	84.93 (13.98)	Test 1	83.09 (16.68)
		Test 2	86.34 (10.73)
		Test 3	85.37 (14.16)
None ($n = 20$)	81.06 (15.05)	Test 1	80.12 (16.26)
		Test 2	81.34 (16.49)
		Test 3	81.71 (12.87)

[a]No significant main effect of repeated tests, $F(2,134) = 1.02$, $p = .36$, $\eta_p^2 = .02$.
[b]Significant main effect of simulation experience, $F(2,61) = 4.02$, $p = .02$, $\eta_p^2 = .12$.
[c]No significant interaction of simulation experience and repeated tests, $F(4,122) < 1$, $p = .98$, $\eta_p^2 = .003$; no significant main effect of repeated tests (means not shown), $F(2,122) = 1.31$, $p = .27$, $\eta_p^2 = .02$.

two effects ($\eta_p^2 = .02$ and $\eta_p^2 = .003$, respectively, in Table 5) make it clear that there was no practice effect; performance did not improve across tests regardless of the level of previous simulation experience.

There was a significant main effect of simulation experience, however, with the best performance occurring in the subgroup with frequent simulation experience and the worst performance occurring in the subgroup with no simulation experience (see Table 5). Bonferroni corrected pairwise comparisons indicated that the subgroup with frequent experience scored 13.6% higher (11.04 points), $p = .02$, than the subgroup with no experience. The subgroup with occasional experience, however, was not significantly different from either the frequent or the no experience subgroups, $p > .15$. Whether the advantage of simulation experience stems from traits of nurses who seek out simulation experience or from recent educational or work experiences that included frequent simulation cannot be determined in this project.

Evidence Regarding Relationships With Conceptually Related Constructs: Self-Assessment

More than 90% of participants completed a self-assessment form. Self-Assessment Scores ranged from 54% to 100%, $n = 61$, $M = 87.63$, $SD = 8.97$. Overall Scores for this subset of participants ranged from 12% to 95%, $n = 61$, $M = 71.73$, $SD = 17.88$. The bivariate correlation showed no meaningful relationship between Self-Assessment Scores and Overall Scores: Pearson $r = .07$. Despite the similarity of the items on assessment form and the NPP rating instrument, these two sources of information about competency (self vs. observer) appear to be independent of each other.

Evidence Regarding Relationship With a Criterion: Supervisor Assessment

Supervisor assessment forms were returned by only 28% of participants' supervisors. Supervisors consistently rated participants very highly, $n = 19$, $M = 94.96$, $SD = 7.87$. Overall Scores for this subset of participants ranged from 49% to 98%, $n = 19$, $M = 76.77$, $SD = 13.14$. The bivariate correlation indicated a small, positive correlation between the two variables: Pearson $r = .31$. Although this result corresponds to the expected relationship between a proposed measure of competency and an existing criterion measure, care should be taken when interpreting the correlation. The small sample size and preponderance of high ratings suggests that supervisor assessment outside the context of workplace requirements is not an appropriate criterion variable for evaluating the validity of simulation-based competency testing.

Evidence for Intended Uses and Interpretations: Performance Subgroups

One of the purposes for developing overall scores was to be able to categorize participants (i.e., to distinguish between low and high competency) and explore the relationship of overall performance to specific patterns of practice breakdown. The Visual NPPs of the performance subgroups were quite different (Figure 2). Participants in the high-performance subgroup rarely failed an item. Those in the moderate-performance group occasionally failed items on one test. Those in the low-performance group frequently failed items on multiple tests. Figure 2 shows that large differences among the low-,

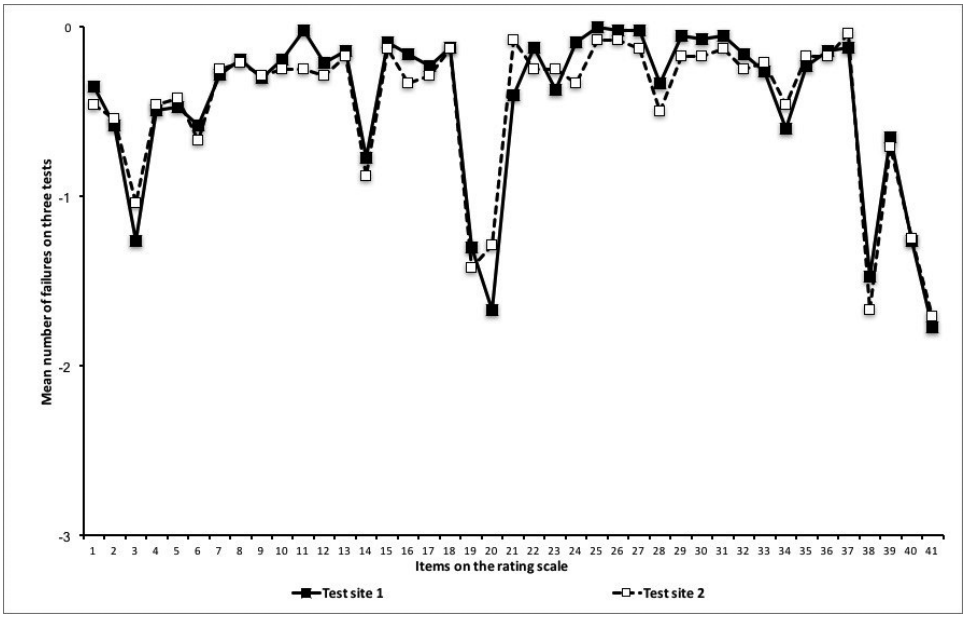

Figure 2. Average visual nursing performance profiles for performance subgroups.

moderate-, and high-performance subgroups existed on specific competencies that are commonly cited as reasons for reporting nurses to the BON (Randolph & Ridenour, 2015). Differences on three representative items are summarized in Table 6. None of the nurses in the low-performance subgroup passed the item "Administers medications accurately and safely" on all three tests, and less than half in the moderate-performance group did, but most nurses in the high-performance group did. A similar pattern was seen on the item "Documents accurate and legally defensible account of interventions" and on the item "Demonstrates understanding of implications of medications and/or interventions." This relationship between overall performance and well-known challenges to safe practice is evidence for construct validity (IOM, 2007, 2011) and supports the use of overall scores as a quick screen for level of performance.

TABLE 6. Competence Differences Among Performance Subgroups

		Overall Score	Medication Administration		Documentation		Critical Thinking	
Subgroup	*n*	*M* (*SD*)	Pass	Fail[a]	Pass	Fail[a]	Pass	Fail[a]
High performance	12	91.06 (3.34)	75%	25%	75%	25%	100%	0%
Moderate performance	42	75.90 (7.05)	36%	64%	19%	81%	88%	12%
Low performance	13	44.65 (17.31)	0%	100%	0%	100%	39%	61%

[a]Passed on all three tests; failed on one or more tests.

TABLE 7. Demographic Characteristics That Distinguish Performance Subgroups

		Age 50+		Highest Education		Simulation Experience	
Subgroup	*n*	Yes	No	ADN	BSN+	None[a]	Some[b]
High performance	12	33%	67%	8%	92%	9%	91%
Moderate performance	42	54%	46%	14%	86%	28%	72%
Low performance	13	77%	23%	31%	69%	62%	38%

Note. ADN = associate's degree in nursing; BSN+ = bachelor's degree in nursing or higher.
[a]"No" simulation experience.
[b]"Occasional" or "frequent" simulation experience.

Performance subgroups also allowed us to identify demographic and professional attributes that may be related to practice breakdown. As summarized in Table 7, nurses in the low-performance subgroup were more likely to be older than 50 years, have an associate's degree as the highest level of education, and have no simulation experience. This relationship between overall performance and age, educational level, and simulation experience is consistent with recent reviews of factors associated with good patient outcomes (Benner et al., 2010; IOM, 2011) and is evidence of construct validity.

Reliability/Generalizability Coefficients: Inter-Rater Reliability

Given that raters were using all four possible ratings, the probability that two or three raters would agree by chance was 40/64 = 0.625. A binomial test was used to compare the actual level of agreement across three raters to the chance level (testing the null hypothesis that observed agreement was 0.625). The test was repeated for each of the 41 items for Test 1, Test 2, and Test 3 (with Bonferroni correction). For example, on Item 1 of Test 1, two or three raters agreed for each participant—across 67 participants, raters agreed 100% of the time, significantly more often than would be expected by chance alone, $p < .00001$. Similarly, on Item 2 of Test 1, two or three raters agreed on the ratings for 65 of 67 participants, 97% agreement, $p < .00001$. This pattern was seen on all items on all tests: On Test 1, the level of agreement among raters ranged from 96% to 100% across the 41 items, $p < .00001$ for every item; on Test 2, it ranged from 93% to 100%, $p < .00001$; and on Test 3, it ranged from 96% to 100%, $p < .00001$.

Reliability/Generalizability Coefficients: Internal Consistency

The Cronbach's alpha was .90 for the Pass–Fail Score and .93 for the Profile Score ($N = 67$).

Reliability/Generalizability Coefficients: Test–Retest Reliability

Participants were tested three times in a single day to give them multiple opportunities to demonstrate their ability to provide safe care in a basic medical-surgical nursing scenario. The three Test Scores were strongly and positively intercorrelated: $r_{1,2} = .61$; $r_{1,3} = .57$; and $r_{2,3} = .69$. A participant's score on one test accounted for 33%–48% of the variance in another, indicating adequate test–retest reliability. These intercorrelations, combined with the lack of practice effects, indicate consistency of measurement across repeated tests.

TABLE 8. Performance Differences Across Testing Sites

	Between-Groups *t* Test	
	n	*M* (*SD*)[a]
Site 1	43	72.83 (16.94)
Site 2	24	72.05 (19.10)

[a]No significant difference between sites, $t(65) < 1$, $p = .86$, $\eta_p^2 < .001$.

Factors Affecting Reliability/Precision: Reproducibility/Standardization

A critical component of reliability/precision is demonstrating that testing procedures can be standardized across testing sites and versions of the test. An independent samples *t* test of Overall Scores showed no significant difference between groups tested at the two sites (Table 8). A plot of the Profile Score comparing performance of the participants who took the test at different sites shows no meaningful difference between sites (Figure 3). A between-groups one-way ANOVA of Overall Scores indicated that performance on the three sets of the test was not significantly different (Table 9). A plot of the Profile Score comparing sets of the test shows no meaningful differences between sets (Figure 4). The very small effect sizes found in these two analyses and seen in the Visual NPPs are evidence of reproducibility/standardization across testing sites and versions of the test.

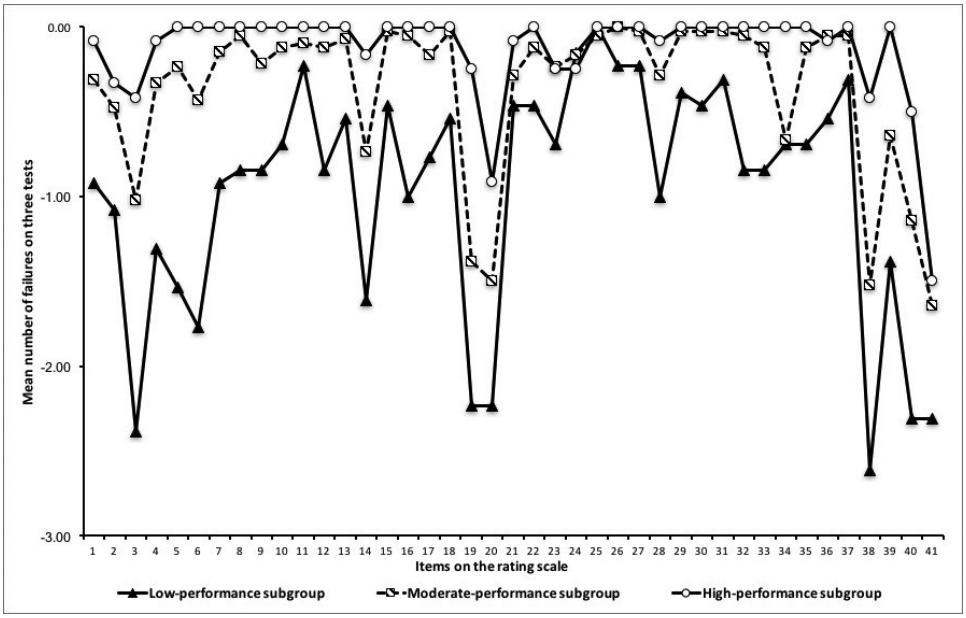

Figure 3. Average visual nursing performance profiles for testing locations.

TABLE 9. Performance Differences Across Test Sets

	One-Way Between-Groups Analysis of Variance	
	n	*M* (*SD*)[a]
Set 1	23	75.72 (12.21)
Set 2	23	70.20 (18.78)
Set 3	21	71.66 (21.27)

[a]No significant main effect of sets, $F(2,64) < 1$, $p = .55$, $\eta_p^2 = .02$.

DISCUSSION

Results from this project will be discussed in three sections. First, the discussion of the instrument will (a) provide evidence for the validity and reliability/precision of the NPP and (b) support the credibility of using the NPP to identify areas for which individuals need additional education or training to promote safe practice. Second, a discussion of nursing performance data will (a) indicate that NPP scores are not a simple proxy for education level, simulation experience, or years of nursing experience and (b) suggest that NPP measures of competency have negligible relationship to self-assessment and supervisor assessment surveys. Last, the discussion will establish how the NPP process represents a new standard for assessing competence and supporting decisions intended to protect public safety.

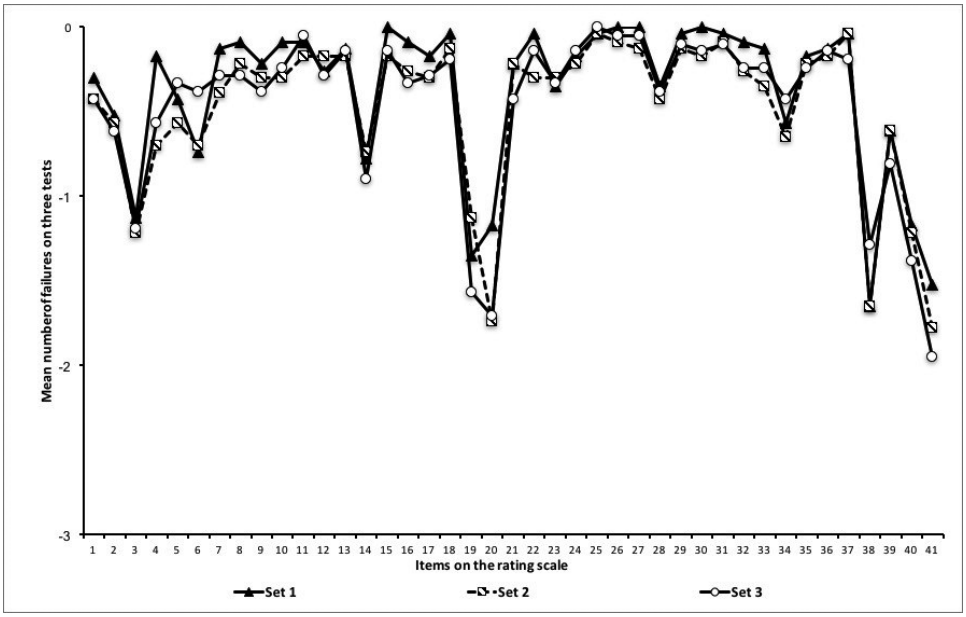

Figure 4. Average visual nursing performance profiles for test sets.

The Nursing Performance Profile Instrument

The evidence that the NPP yields highly reproducible results is especially critical to the BON, whose decisions must be legally defensible and socially just. One of the most promising findings of this project is the high degree of standardization achieved. Variability because of repeated testing, sets of testing scenarios, and testing locations was small, accounting for less than 1% of the variance in NPP scores. These results suggest that the NPP simulation testing processes and procedures could be used by other schools and regulatory agencies to assess nursing competence.

Of particular interest to regulators, educators, and employers is the evidence that the NPP was sensitive to common errors associated with significant risk to patients. Items frequently failed during NPP simulation tests are consistent with nursing practice difficulties identified in the literature related to medication errors, infection control, documentation, and telephone orders (Benner et al., 2010; Bensfield et al., 2012; IOM, 2011), supporting the credibility of the instrument and the testing process.

Medication administration errors in hospitals are prevalent with slips and lapses identified as the most common contributing factors (Keers, Williams, Cooke, & Ashcroft, 2013). There were a variety of medication errors demonstrated during the NPP simulation tests. A common error was administering the wrong volume and therefore wrong dose of an intravenous medication despite having the correct volume listed on the medication administration record. Of further concern, participants in this study who made a medication error almost never self-identified the error, which is consistent with previous reports that medication errors are more often detected as a result of direct observation rather than through self-reporting by nurses (Buckley, Erstad, Kopp, Theodorou, & Priestley, 2007).

Participants in the NPP project frequently failed to demonstrate basic infection control principles such as consistent hand hygiene. Magill et al. (2014) found that on a daily basis, approximately 1 of every 25 patients in U.S. hospitals has at least one health care–associated infection. Out of 968 observations of 123 health care workers, 23.2% complied with the 2009 World Health Organization recommended moments for hand hygiene while self-reporting 82.4% compliance (Eiamsitrakoon, Apisarnthanarak, Nuallaong, Khawcharoenporn, & Mundy, 2013).

Many NPP participants had difficulty documenting accurate and legally defensible accounts of the care they provided during the testing scenarios. Documentation errors ranged from omission of critical information to inaccurate data, including listing actions that the nurse was not observed performing. Some participants used unconventional abbreviations or other idiosyncratic documentation methods that were difficult to decipher. These documentation problems are consistent with the analysis of 341 hospital records from 10 hospitals in the Netherlands that revealed nursing documentation was often incoherent, inaccurate, and contained easily misunderstood nonstandard abbreviations (Paans, Sermeus, Nieweg, & van der Schans, 2010). Telephone orders that occurred during NPP testing scenarios were often recorded in a manner inconsistent with what the physician actor actually said and were sometimes recorded in nurse's notes rather than in the physician's orders. It has long been recommended that verbal orders be avoided to reduce miscommunications linked to medication errors (Cohen, 2000).

Nursing Performance

The NPP scores combined with the raters' comments about participant performance suggest that competence is the result of a complex interaction of education, length of nursing

experience, and experience with simulation. First, nursing practice has evolved over the past 25 years as more evidence has been integrated into practice standards (Beyea & Slattery, 2013). Minimally safe performance behaviors assessed by the NPP, such as using two patient identifiers, were not part of standard nursing practice 25 years ago and so may be effortful for older nurses to incorporate into their practices. Second, nursing education/testing technologies, including simulation with high-fidelity manikins or standardized patients, have been available for two decades but have become commonplace more recently. Thus, "younger" nurses, who are recent graduates of basic or advanced educational programs and nurses certifying in critical care settings may have had more opportunities to incorporate simulation and related technologies into practice. Third, independent of simulation experience, higher education was associated with safer practice in our testing. This finding is congruent with a persuasive body of evidence that patient outcomes are improved when staffing policies are changed to increase the number of nurses with bachelor's degrees (Rosseter, 2015).

The NPP scores were not well correlated with self-assessment and supervisor assessment surveys. M. J. Gordon's (1991) review of the literature indicates that self-assessment by health professions students is largely unrelated to and unaffected by experts' ratings or objective tests. Although much has been written about the usefulness of "360 degree performance review" or "multisource assessment," which includes self, peer, subordinate, and supervisor assessments (Smither, London, & Reilly, 2005), this form of assessment has not been well studied in nursing (Garbett, Hardy, Manley, Titchen, & McCormack, 2007). The typical finding across a broad range of disciplines is that supervisor assessment has a modest correlation with self-assessment or peer assessment and so these multiple sources of feedback are seen as representing independent perspectives rather than converging evidence of competence/incompetence.

Although our analysis of the relationship between supervisor and rater ratings was hampered by a low return rate for supervisor assessments, the small correlation between supervisor and rater ratings may reflect the raters' opportunity to identify errors from direct observation of the recorded performances. The supervisor may be relying on indirect sources of data such as general impressions of the nurse, patient satisfaction surveys, or the reports of others to form an assessment of the nurse's competency. In contrast, using a process such as the NPP has the potential to alert managers to unsafe nursing behaviors before they cause actual harm to the patient or health care institution.

A New Standard and the Future of Assessing Nursing Competence

Missing from the literature, until now, is a multiple-perspective comparison of competence ratings based on self-assessment, supervisor assessment, and blinded rater observations of nurses' performance. Connection of these three perspectives in the NPP project provides a clearer view of the limitations of self-assessment and supervisor assessment that do not rely on direct observation. Future projects should have nurses and their supervisors rate the nurses' videotaped performances to determine how well the three perspectives of self, supervisor, and rater match when all are using direct observation as the basis for the assessment.

Because of the small and unequal participant groups with differing characteristics (i.e., education, practice area, years of experience, and/or simulation experience), further study is needed to compare NPP scores to factors traditionally linked to continued nursing competence. A future project could include intentionally selecting participants to represent diverse demographics, educational levels, and extent of simulation experience.

Additional exploration is needed to determine how the NPP process could be used to further support patient safety, education/remediation, continued competence, practice regulation (IOM, 2011), and transfer of simulation learning to actual patient care (Hughes, 2008). Use of the NPP process to investigate how nursing practice behaviors are influenced by system factors (medication packaging or dispensing processes) could be explored within the context of patient management scenarios. Health care systems may want to use such a process to investigate the return on investment from implementing patient safety initiatives. Educators in prelicensure nursing programs may be able to use a process similar to the NPP for formative assessment and summative evaluation of their students nearing graduation. Employers may be interested in exploring the use of simulation performance testing to evaluate the efficacy of orientation training for new employees or remediation strategies for nurses with practice breakdown behaviors. Regulators in other states and countries may be able to adapt the NPP process to support regulation of nursing competence in their jurisdictions. Thus far, the NPP process has been applied to the common nursing work setting of adult acute care. Further development of testing scenarios for diverse patient populations and work settings could provide broader opportunities to study transfer of knowledge from simulation to performance in health care settings and patient outcomes.

CONCLUSIONS

Project results demonstrate how a snapshot of a nurse's practice can be created by recording behavior during realistic simulation scenarios. Results demonstrate the NPP rating instrument can be used in an objective, valid, and reliable/precise manner to identify areas of nursing practice breakdown. The NPP development process described in this article and previous publications (Hinton et al., 2012; Randolph et al., 2012; Randolph & Ridenour, 2015) provides evidence suggesting simulation testing and rating of videotaped performance by trained raters can result in a fair evaluation of a nurse's competence that supports decision-making processes designed to protect patient safety.

REFERENCES

Aebersold, M., & Tschannen, D. (2013). Simulation in nursing practice: The impact on patient care. *The Online Journal of Issues in Nursing*, *18*(2). http://dx.doi.org/10.3912/OJIN.Vol18No02Man06

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

AMN Healthcare. (2013). *2013 Survey of registered nurses*. San Diego, CA: Author. Retrieved from http://www.amnhealthcare.com/uploadedFiles/MainSite/Content/Healthcare_Industry_Insights/Industry_Research/2013_RNSurvey.pdf

Arizona State Board of Nursing. (2007). *Continued competence. Information reports*. Phoenix, AZ: Author.

Benner, P., Malloch, K., Sheets, V., Bitz, K., Emrich, L., Thomas, M. B., . . . Farrell, M. (2006). TERCAP: Creating a national database on nursing errors. *Harvard Health Policy Review*, *7*(1), 48–63.

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

Bensfield, L. A., Olech, M. J., & Horsley, T. L. (2012). Simulation for high-stakes evaluation in nursing. *Nurse Educator*, *37*(2), 71–74.

Beyea, S. C., & Slattery, M. J. (2013). Historical perspectives on evidence-based nursing. *Nursing Science Quarterly*, *26*(2), 152–155. http://dx.doi.org/10.1177/0894318413477140

Buckley, M. S., Erstad, B. L., Kopp, B. J., Theodorou, A. A., & Priestley, G. (2007). Direct observation approach for detecting medication errors and adverse drug events in a pediatric intensive care unit. *Pediatric Critical Care Medicine*, *8*(2), 145–152. http://dx.doi.org/10.109/01.PCC.0000257038.39434.04

Budden, J., Zhong, E., Moulton, P., & Cimiotti, J. (2013). Highlights of the national workforce survey of registered nurses. *Journal of Nursing Regulation*, *4*(2), 5–14.

Cohen, M. R. (2000). *Medication errors: Causes, prevention, and risk management*. Boston, MA: Jones & Bartlett Learning.

Davis, D. A., Mazmanian, P. E., Fordis, M., Van Harrison, R., Thorpe, K. E., & Perrier, L. (2006). Accuracy of physician self-assessment compared with observed measures of competence: A systematic review. *Journal of the American Medical Association*, *296*, 1094–1102.

Department for Professional Employees. (2015). *Nursing: A profile of the profession*. Washington, DC: Author. Retrieved from http://dpeaflcio.org/programs-publications/issue-fact-sheets/nursing-a-profile-of-the-profession/

Dreyfus, S. E., & Dreyfus, H. L. (1980). *A five-stage model of mental activities involved in directed skill acquisition.* Unpublished report, University of California, Berkeley, CA.

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest*, *5*(3), 69–106.

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes*, *105*(1), 98–121.

Eiamsitrakoon, T., Apisarnthanarak, A., Nuallaong, W., Khawcharoenporn, T., & Mundy, L. M. (2013). Hand hygiene behavior: Translating behavioral research into infection control practice. *Infection Control and Hospital Epidemiology*, *34*(11), 1137–1145. http://dx.doi.org/10.1086/673446

Foronda, C., Liu, S., & Bauman, E. B. (2013). Evaluation of simulation in undergraduate nurse education: An integrative review. *Clinical Simulation in Nursing*, *9*(10), e409–e416. http://dx.doi.org/10.1016/j.ecns.2012.11.003

Garbett, R., Hardy, S., Manley, K., Titchen, A., & McCormack, B. (2007). Developing a qualitative approach to 360-degree feedback to aid understanding and development of clinical expertise. *Journal of Nursing Management*, *15*(3), 342–347. http://dx.doi.org/10.1111/j.1365-2834.2007.00692.x

Garside, J. R., & Nhemachena, J. Z. Z. (2013). A concept analysis of competence and its transition in nursing. *Nurse Education Today*, *33*, 541–545.

Gordon, C. J., Frotjold, A., & Bloomfield, J. G. (2015). Nursing students' blood pressure measurement accuracy during clinical practice. *Journal of Nursing Education and Practice*, *5*(5), 46–54.

Gordon, M. J. (1991). A review of the validity and accuracy of self-assessments in health professions training. *Academic Medicine*, *66*(12), 762–769.

Hinton, J. E., Mays, M. Z., Hagler, D., Randolph, P., Brooks, R., DeFalco, N., . . . Weberg, D. (2012). Measuring post-licensure competence with simulation: The nursing performance profile. *Journal of Nursing Regulation*, *3*(2), 45–53.

Hughes, R. G. (Ed.) (2008). *Patient safety and quality: An evidence-based handbook for nurses*. (AHRQ Publication No. 08-0043). Rockville, MD: Agency for Healthcare Research and Quality.

Institute of Medicine. (2007). *Preventing medication errors*. Washington, DC: The National Academies Press.

Institute of Medicine. (2011). *The future of nursing: Leading change, advancing health*. Washington, DC: The National Academies Press.

Johnson, W. G., Wilson, B., Edge, M., Qiu, Y., Oliver, E. L., & Russell, K. (2009). *The Arizona healthcare workforce: Nurses, pharmacists, & physician assistants*. Tempe, AZ: The Center for Health Information & Research.

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

Keers, R. N., Williams, S. D., Cooke, J., & Ashcroft, D. M. (2013). Causes of medication administration errors in hospitals: A systematic review of quantitative and qualitative evidence. *Drug Safety*, *36*, 1045–1067. http://dx.doi.org/10.1007/s40264-13-0090-2

Magill, S. S., Edwards, J. R., Bamberg, W., Beldavs, Z. G., Dumyati, G., Kainer, M. A., . . . Fridskin, S. K. (2014). Multistate point-prevalence survey of health care-associated infections. *The New England Journal of Medicine*, *370*(13), 1198–1208.

Meakim, C., Boese, T., Decker, S., Franklin, A. E., Gloe, D., Lioce, L., . . . Borum, J. C. (2013). Standards of best practices: Simulation. Standard I: Terminology. *Clinical Simulation in Nursing*, *9*(65), S3–S11. http://dx.doi.org/10.1016j.ecns.2013.04.001

National Council of State Boards of Nursing. (2007). Li's clinical competency assessment of newly licensed nurses. In *The Impact of Transition Experience on Practice of Newly Licensed Registered Nurses*. Business Book, NCSBN Annual Meeting: Navigating the Evolution of Nursing Regulation. Chicago, IL: Author.

Numminen, O., Leino-Kilpi, H., Isoaho, H., & Meretoja, R. (2015). Congruence between nurse managers' and nurses' competence assessments: A correlation study. *Journal of Nursing Education and Practice*, *5*(1), 142–150.

Paans, W., Sermeus, W., Nieweg, R. M. B., & van der Schans, C. P. (2010). D-catch instrument: Development and psychometric testing of a measurement instrument for nursing documentation in hospitals. *Journal of Advanced Nursing*, *66*(6), 1388–1400. http://dx.doi.org/10.1111/j.1365-2648.2010.05302.x

Quality and Safety Education for Nurses. (2010). *Competency KSAs*. Retrieved from http://www.qsen.org/competencies/graduate-ksas/#patient-centered_ care

Randolph, P. (2016). *Arizona fact sheet*. Phoenix, AZ: Arizona Action Coalition. Retrieved from http://www.futureofnursingaz.com/home/arizona-fact-sheet/

Randolph, P. K., Hinton, J. E., Hagler, D., Mays, M. Z., Kastenbaum, B., Brooks, R., . . . Weberg, D. (2012). Measuring competence: Collaboration for safety. *Journal of Continuing Education in Nursing*, *43*(12), 541–549.

Randolph, P. K., & Ridenour, J. (2015). Comparing simulated nursing performance to actual practice. *Journal of Nursing Regulation*, *6*(1), 33–38.

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. http://dx.doi.org/10.5480/15-1639

Rosseter, R. J. (2014). *Nursing shortage fact sheet*. Washington, DC: American Association of Colleges of Nursing.

Rosseter, R. J. (2015). *Creating a more highly qualified nursing workforce*. Washington, DC: American Association of Colleges of Nursing.

Smither, J. W., London, M., & Reilly, R. R. (2005). Does performance improve following multisource feedback? A theoretical model, meta-analysis, and review of empirical findings. *Personnel Psychology*, *58*(1), 33–66. http://dx.doi.org/10.1111/j.1744-6570.2005.514_1.x

Whyte, J., Pickett-Hauber, R., Ward, P., Eccles, D. W., & Harris, K. R. (2013). The relationship between standardized test scores and clinical performance. *Clinical Simulation in Nursing*, *9*(12), e563–e570. http://dx.doi.org/10.1016/j.ecns.2013.05.006

Williams, R. G., Klamen, D. A., & McGaghie, W. C. (2003). Cognitive, social and environmental sources of bias in clinical performance ratings. *Teaching and Learning in Medicine*, *15*(4), 270–292.

Acknowledgments. Funding for the project was received from the National Council of State Boards of Nursing Center for Regulatory Excellence.

Correspondence regarding this article should be directed to Janine E. Hinton, PhD, RN, Scottsdale Community College, 9000 E. Chaparral Rd, Scottsdale, AZ 85256. E-mail: Janine.Hinton@scottsdalecc.edu