ISSN 2475-529X

Research Article

Journal of Nursing & Healthcare

Development of Quantum, an Instructor-Mediated Performance Assessment Test, and Student Measure Validation

Stephen C. Hetherman[1]*, Lori Lioce[2], Lucille Gambardella[3] and Bonita L. Longo[4]

[1]*Objectivity Plus.*

[2]*Clinical Associate Professor, Executive Director Learning and Technology Resource Center The University of Alabama in Huntsville College of Nursing, USA.*

[3]*Professor Emerita and Chair Wesley College CEO, Positive Transitions.*

[4]*Nurse Educator Consultant*

***Corresponding author**
Stephen C. Hetherman, Objectivity Plus, 3909 Ambassador Caffery Pkwy., Bldg. K, Lafayette, LA 70503, USA, E-mail: steve@objectivityplus.com.

Submitted: 14 July 2017; **Accepted**: 21 July 2017; **Published**: 25 July 2017

Abstract

This article discusses Quantum, an instructor-mediated performance assessment test, used in simulation to assist nurse educators to objectively measure student performance and document clinical competency. The article describes the process of development, validation, and measurement of student's integrated knowledge, skills, attitudes, and clinical reasoning used in decision-making to improve safe effective nursing practice. The article provides a review of Quantum's evaluation framework including: patient safety, assessment, communication, intervention, and documentation. The sample comprised of pre-licensure student nurses enrolled in core nursing courses from 14 nursing programs. Classical and Rasch data analyses found empirical evidence in support of the reliability of measures and validity of inferences in terms of content validity, optimum scoring structure, unidimensionality, reliability, invariance, responsiveness, consequential validity, and interpretability.

Introduction

Tests of knowledge by ATI, HESI, and Kaplan are important, but they portray an incomplete student appraisal if we really believe there is more to the practice of nursing than knowing. According to research published by Miller in 1990 the cognition zone ('knows' or 'knows how') correlates poorly with the behavior zone ('shows how' or 'does') [1]. Benner, Sutphen, Leonard, and Day (2010), in their book *Educating nurses: A call for radical transformation*, recommended varying the use of assessment of student performance beyond an exclusive focus on multiple-choice exams [2]. In 2014, the National Council of State Boards of Nursing's National Simulation Study (NCSBN NSS) provided substantial evidence that up to 50% simulation can be effectively substituted for traditional clinical experience in pre-licensure core nursing courses [3]. These recommendations coupled with diminishing clinical sites, high student-faculty ratios, and the need to evaluate multiple learning domains [4], may prompt nursing programs to use a performance assessment test in their simulation laboratory to replace a portion of traditional clinical evaluation. However, validity and reliability evidence has been lacking on instruments to assist nurse educators with being more objective [5]. The subjective nature of clinical evaluation in education and practice stems from the inclusion of instructors in the performance assessment, which complicates the measurement process. Consequently, there are no psychometrically-sound performance assessment tests to objectively measure a student's performance in simulation [6, 7]. Therefore, documentation was needed on the psychometric qualities of a performance assessment test.

Purpose

An instructor-mediated, performance assessment test designed to objectively measure a student's clinical competence presents a number of problems, namely, modelling and accounting for sources of error variance that undermine student test score interpretation. There has been little effort in this area to collect and integrated various sources of evidence and theory to support the intended interpretation of test scores derived from an instructor-mediated, performance assessment test. The purpose of this article was to execute an empirical validation investigation of an instructor-mediated, performance assessment test to resolve these problems. To achieve this purpose, this validation study used Quantum as the instructor-mediated, performance assessment test in 14 nursing programs' simulation laboratories.

Background

Objectively evaluating students' clinical competency can be the most daunting duty of an educator, whether it is in the clinical setting or the simulation laboratory [8]. As such, the development of an instructor-mediated, performance assessment test to measure a student's clinical competence plays a large role since it is a time-consuming and difficult process that requires multiple skill sets and conditions [9] including psychometric and subject matter expertise and adherence to measurement best practices [10]. The test development process typically involves two or three iterations of data collection and item revision in conjunction with expert reviews [11]. Measuring complex learning outcomes across all three domains of learning has historically been challenging, but Li (2007) and Kardong-Edgren, et al., (2010) outlined how simulation offers opportunities to evaluate them collectively [5, 12]. Simulation has become an environment to increase student engagement and learning outcomes, but, researchers have shown student-learning outcomes have not been measured in a reliable and valid manner [13, 14].

Student learning outcomes in simulation is attracting a lot of attention in higher education due to the convergence of several factors including: simulation costs associated with student success, retention rates, and learning gains; accreditation standards; increasing efforts to link higher education funding to student success; and NCSBN NSS's findings [3]. Outcome assessment should address both the student success and learning gains that result from simulation [15]. "Soft" benefits, such as improved decision-making including the ability to solve clinical problems by setting patient care priorities or responding to a change in the patient's condition [2], can be further analyzed to trace through to "hard" benefits (e.g., improvements in measures such as student retention). The Joint Commission on Accreditation of Healthcare Organizations (JCAHO) noted that healthcare organizations have historically relied on education and experience to support competence, but an increasing number of healthcare organizations are seeking objective measures of a nurse's knowledge that is required for safe practice [16]. The anticipated benefits of using an objective performance assessment's data are one, you tend to improve what you measure; and two, you simply can't manage what you don't. Whether it is to drive an educational intervention to improve safe practice or safely decrease the length of orientation for new employees using simulation for teaching and learning reflects a presumption that learning is taking place and the teaching is effective. How do you know if you don't measure it objectively?

Method

Performance assessments offered by the simulation industry or espoused in the literature to measure a student's clinical competency do not model all variables of the performance assessment and do not account for their effects on students' test scores. For example, the difficulty of the task performed and characteristics of instructors (e.g., the severity of particular instructor, their consistency, the way they interpret the rating guidelines) are crucial in determining the pattern of scores allocated to students in a performance task, and these sources of variation in a performance assessment must be modelled in order to provide fair and objective student test scores. The aim of this current study was to develop an instructor-mediated, performance assessment test, Quantum, to examine possible sources of error that reduce the validity of students' test scores.

Test Development Design

Quantum was developed by the authors and subject matter experts (SMEs) who understand how important it is to: remove subjectivity from the rating process; have reliable student measures; quantify a student's performance; and document clinical competency. A psychometrician and SMEs oversaw each clinical scenario's test blueprint from student preparation, learning objectives, scenario content, simulation set-up document, evaluation criteria, rating guidelines, feedback, and debriefing [17].

Each test blueprint was linked to several educational learning theories: *talk aloud protocol* [18], the *generation effect* [19], *learning progression* [20, 21], and *reflective practice* [22-24]; applicable standards [15, 25-27]; references [28, 29]; current content [30]; legal aspects of clinical nursing education [31]; and reflects the framework for clinical assessment and the Nursing Skill Development and Clinical Judgment Model [1, 17].

Quantum was first developed for registered nurse (RN) programs and then for license practical (vocational) nurse (LPN/LVN) programs. Quantum-RN evaluated a student nurse's performance across five subdomains: patient safety, assessment, communication, intervention (including I-SBAR-R (Grbach, n.d.)) [32], and documentation. Quantum-LPN/LVN evaluated a student nurse's performance across the same subdomains, but data collection replaced assessment.

Quantum was comprised of several standardized clinical scenarios that each student must successfully perform in order to demonstrate clinical competence. Each clinical scenario has specific descriptions of activities that demonstrate the performance at basic, intermediate and advanced levels.

Theoretical Framework for Test Development

Quantum was developed using Chatterji's Process Model [33]. The Process Model allowed the researchers to merge the issues of logical analyses of test content and empirical confirmation of the variables in Quantum, both essential to defending the validity of test score interpretations [10]. These iterative procedures employed for validation combined both classical test theory (CTT) and Rasch analyses to document the psychometric qualities of Quantum [34, 35].

Validation Study Sample

In 2012, 137 pre-licensure nursing programs throughout the United States were notified of this validation study and its requirements via email. The invitation was open to Board of Nursing approved LPN/LVN, ADN, and BSN nursing programs who use simulation in any one of its core pre-licensure nursing courses. Interested nursing programs had to have access to a simulation laboratory, and a willingness to designate a point of contact of faculty, and support staff. All types of schools and program levels with comparable clinical course curricula, and demographic data that represented a diverse student population were selected. One specific criterion was for students enrolled in, or who just completed a core pre-licensure nursing course. After a field-test participation conference call with interested academic stakeholders, 14 nursing programs participated from geographically diverse areas, namely the South (71%), Northeast (7%), West (14%), and Midwest (7%) as shown in Table 1.

	BSN(4)	ADN(5)	LPN/LVN(5)
Health Assessment	116	126	0
Fundmentals	109	140	117
Medical/Surgical	105	111	135
Obstetrics	102	108	140
Pediatrics	103	105	131
Mental Health	101	107	116
Leadership	93	32	0

Table 1: Frequency Distribution of Sample by Core Pre-Licensure Nursing Course.

Participating nursing programs volunteered to assist in the study. Approval of the institutional review board at each institution was obtained prior to obtaining faculty consent and commencing data collection. During the pre-briefing sessions, all student nurses who consented to participate in the study were reminded that they could opt-out up to two weeks before the field test. The purpose of the field test process was to evaluate the psychometric properties of the data arising from Quantum.

Procedure

Faculty training, simulation set-up, and staff support prior to the administration of the field test were provided to each school to ensure standardization. To maximize scoring consistency, all instructors involved were trained in the use of Quantum's touch screen Tablet technology and were given access to the on-line tutorial to reinforce their training. Instructors not present at rater training were provided with an on-line tutorial to access before the first test administration. Subsequent rater training webinars were scheduled as requested. Faculty found the Tablet intuitive and became more comfortable with each use. Equivalence estimates reflect a high degree of inter-rater agreement as detailed under the heading, *Reliability*, and shown in Table 3, column labelled ICC.

Student nurses who volunteered were assigned a time and date to perform the associated clinical scenario. Two instructors, serving as raters, used two Android Tablets and logged into their previously created accounts. Their test administration schedules were listed on the Tablet.

At the designated test time, a student arrived and completed the security protocol by verifying his or her identity with a picture, signature, and voice pattern. Each student was given a ten-minute preparation period to read the in-App scenario of the patient's situation to be managed, and took notes on flow sheets. The preparation period was followed by a 20-minute performance period when both instructors assessed the student's performance independently using the evaluation criteria on the in-App rating sheet specific to the scenario being tested. At the end of the performance period, the student completed a post-simulation attitudinal survey regarding his or her simulation experience.

Data collection

Data collection strictly followed the ethical standards of the participants' schools. Data collection combined with field tests is called a validation study [11]. Field test schedules used the laws of parsimony to ensure data connectivity.

Data analysis

CTT approaches [36] and Rasch measurement models [37, 38] were applied to the data using SPSS version 16 [39], Winsteps (version 3.55) and Facets (version 3.47) programs to examine a number of validity aspects of Quantum [40, 41]. These measurement approaches were used to verify that the evaluation criteria making up each subdomain in a clinical scenario were underpinned by a single construct; whether the rating scale functions well; whether there was good targeting between the evaluation criteria and student abilities; whether clinical scenarios were sensitive enough to detect changes in student test scores following intervention; whether major sources of error variance were addressed; whether the evaluation criteria met the measurement criteria of invariance for student test scores; whether the value implications of test score interpretations as a source of action; and to what degree can qualitative meaning be assigned to quantitative measures. Only the core evaluation criteria common to all RN and LPN/LVN participating programs were analyzed in this validation study.

Results

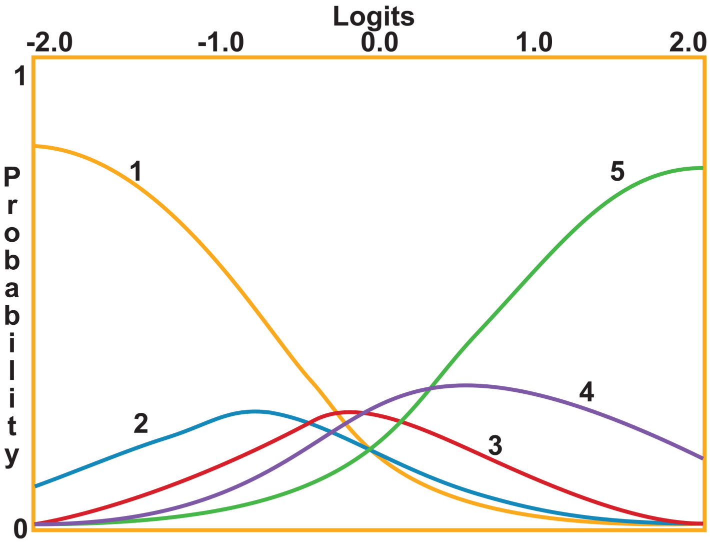
The validity of an intended interpretation of test scores relies on all available evidence [10]. As such, rigorous evaluation statistics regarding evidence relevant to the validity aspects of Quantum follow.

Rating Scale Analysis

Quantum relies on the assumption that instructors will evaluate student performances using the particular criteria on the SME designed 5-point scale ranging (1 = *Unacceptable performance* to 5 = *Superior performance*) [10]. Empirical evidence validated a dichotomous scoring structure after subsequent analyses. Consequently, an interval measurement scale and a dependable scoring structure was developed. The interval measurement scale has three distinct features: (1) demonstrates one thing is bigger than the other; (2) determines how much bigger; and (3) lends itself to mathematical computations [42]. These interval measures may then be used in subsequent statistical analyses that assume an interval scale.

In order to address the substantive aspect of validity, empirical evidence shown in Figure 1 and Table 2 relates to the optimal number of rating scale categories for the target construct (i.e., clinical competency).

Figure 1. Probability Curves of 5-Point Rating Scale



Initial scale development assumed that the rating scale was hierarchical, but it failed to produce clear conclusions as depicted in Figure 1, evidence signaling a need to collapse the rating scale structure. In Figure 1, notice at no point was category 2 (*Inconsistent*) the most likely category to be observed. It can be seen that categories 3 (*Effective*) and 4 (*Highly Effective*) represent a narrow range of real performance. Applying Linacre's (2004) guidelines for interpreting Rasch model indices, in column (2) labelled count on Table 2, there were large frequency counts, indicating stable estimates [43]. The shape of each rating scale distribution signaled aberrant rating category usage.

Table 2: Statistics for the Rating Catergories by Clinical Scenario (continued)

Course	LPN/LVN						
(1) Rating Category	**(2) Count**	**(3) %**	**(4) Obs Avg**	**(5) Exp Avg**	**(6) Outfit**	**(7) Step Calib**	**(8) Monotonically Increase**
Fundamentals							
Unsatisfactory	244	17%	-1.28	-1.37	1.00	None	
Inconsistent	187	13%	-0.36	-0.59	1.12	-1.57	
Effective	230	16%	-0.13	0.22	0.73	-0.46	1.11
Highly Effective	244	17%	1.02	1.07	0.95	0.17	0.63
Exceptional	532	37%	3.24	2.15	1.15	1.40	1.23
Medical/Surgical							
Unsatisfactory	218	14%	-1.48	-1.53	1.00	None	
Inconsistent	172	11%	-0.42	0.62	1.25	-2.48	
Effective	374	24%	-0.18	0.37	0.95	-1.05	1.43
Highly Effective	343	22%	1.63	1.59	0.89	-0.66	0.39
Exceptional	452	29%	2.77	2.78	1.00	2.72	3.38
Women's Health							
Unsatisfactory	1626	14%	-1.21	-1.26	1.01	None	
Inconsistent	2787	24%	-0.74	-0.65	1.16	-1.52	
Effective	3368	29%	-0.09	0.09	1.09	-0.52	1.00
Highly Effective	2555	22%	1.83	1.79	1.03	0.24	0.76
Exceptional	1278	11%	2.32	2.34	1.00	2.05	1.81
Pediatrics							
Unsatisfactory	188	16%	-0.95	-1.10	1.01	None	
Inconsistent	270	23%	-0.13	-0.22	1.13	-1.87	
Effective	211	18%	0.06	0.11	0.99	-1.03	0.84
Highly Effective	387	33%	1.40	1.36	0.98	-0.40	0.63
Exceptional	117	10%	2.64	2.65	1.02	1.43	1.83
Mental Health							
Unsatisfactory	227	18%	-1.63	-1.69	1.00	None	
Inconsistent	290	23%	-0.53	-0.45	1.02	-1.43	
Effective	428	34%	0.02	0.07	1.00	-0.61	0.82
Highly Effective	214	17%	1.63	1.67	1.01	0.10	0.71
Exceptional	101	8%	2.71	2.68	1.00	1.92	1.82

As seen in column (4) labelled observed average in Table 2, increasing amounts of student clinical competency corresponds to increasing probabilities of the student being observed in higher rating categories of the rating scale [44]. However, the advances across rating categories were uneven, for instance, in Health Assessment-RN: -1.51 to -0.62 logits (a jump of 2.13), and then from -0.07 to 0.68 (a jump of 0.75). This was indicative of problems with instructors applying the appropriate rating category to the student's performance across evaluation criteria.

A comparison between columns (4) labelled observed average and (5) expected average in Table 2, specifically in the Effective category were contradictory to the intended use of the rating scale in 5 out of 7 RN courses and 3 out of 5 LPN/LVN courses, further evidence signaling a need to collapse this rating category to improve measure stability and accuracy [45].

Column (6) labelled outfit on Table 2 shows all rating categories for each clinical scenario had unweighted mean-squared fit statistics ranging between 0.72 and 1.35 for RN and 0.73 and 1.25 for LPN/LVN. This satisfied Linacre's guideline as the majority of the rating category fit statistics were around 1.0 and all were less than 2.0.

Column (7) labelled step calibration on Table2 shows where the probability curves intersect depicted in Figure 1. As shown, advancement from one-step to the next was between 0.26 and 3.50 logits for RN and 0.39 and 3.38 logits for LPN/LVN, further evidence for combining the rating categories to have wider practical meaning since the advancement was smaller than 0.81 logits for 5-point scale [46].

Table 2: Statistics for the Rating Catergories by Clinical Scenario

Course	RN						
(1) Rating Category	(2) Count	(3) %	(4) Obs Avg	(5) Exp Avg	(6) Outfit	(7) Step Calib	(8) Monotonically Increase
Health Assessment							
Unsatisfactory	959	12%	-1.51	-1.57	1.00	None	
Inconsistent	799	10%	-0.62	-0.53	1.28	-2.26	
Effective	1998	25%	-0.07	0.04	0.92	-0.19	2.07
Highly Effective	1279	16%	0.68	0.57	0.96	0.56	0.75
Exceptional	2957	37%	1.67	1.73	1.01	2.49	1.93
Fundamentals							
Unsatisfactory	1000	13%	-1.31	-1.36	1.00	None	
Inconsistent	846	11%	-0.53	0.45	1.20	-1.68	
Effective	1846	24%	-0.13	0.22	1.03	-0.09	1.59
Highly Effective	1231	16%	0.54	0.44	1.03	0.17	0.26
Exceptional	2769	36%	1.25	1.30	1.00	1.40	1.23
Medical/Surgical							
Unsatisfactory	807	18%	-1.83	-1.59	1.00	None	
Inconsistent	717	16%	-0.60	-1.29	1.20	-2.50	
Effective	986	22%	-0.09	0.09	1.35	-1.03	1.47
Highly Effective	852	19%	1.80	1.79	1.25	-0.50	0.53
Exceptional	1121	25%	3.24	3.22	1.00	3.00	3.50
Women's Health							
Unsatisfactory	1078	19%	-1.16	-1.25	1.00	None	
Inconsistent	737	13%	-0.40	-0.27	1.27	-1.23	
Effective	1305	23%	-0.18	0.37	0.72	-0.41	0.82
Highly Effective	851	15%	0.72	0.67	1.04	0.19	0.60
Exceptional	1702	30%	1.69	1.70	1.03	1.64	1.45
Pediatrics							
Unsatisfactory	273	18%	-0.79	-1.01	0.97	None	
Inconsistent	152	10%	-0.07	0.04	1.12	-1.85	
Effective	288	19%	0.02	0.10	1.06	-0.63	1.22
Highly Effective	258	17%	0.85	0.87	1.01	-0.07	0.56
Exceptional	546	36%	1.85	1.82	1.00	2.48	2.55
Mental Health							
Unsatisfactory	257	18%	-1.67	-1.73	1.00	None	
Inconsistent	200	14%	-0.83	-0.83	1.15	-1.40	
Effective	328	23%	0.00	0.05	0.95	-0.24	1.16
Highly Effective	242	17%	1.06	1.02	0.73	0.17	0.41
Exceptional	399	28%	1.91	1.94	1.04	1.64	1.47
Leadership							
Unsatisfactory	174	19%	-1.21	-1.26	1.00	None	
Inconsistent	138	25%	-0.53	-0.68	1.20	-1.85	
Effective	230	25%	-0.02	0.10	1.18	0.66	1.19
Highly Effective	165	18%	0.85	0.87	1.10	0.09	0.75
Exceptional	211	23%	1.85	1.82	1.03	1.72	1.63

Dimensionality Analysis

Researchers investigated whether or not there was more than one variance component explaining the structure of the performance assessment data. Table 3 showed that the eigenvalues of the second component from the SPSS Factor Analysis routines were less than 1.0 for all clinical scenarios [47, 48]; meaning only one component accounted for considerably more variance than the remaining components. The first factor had maximum variance which ranged from 91.67% for the Medical Surgical-RN Advanced scenario to 76.95% for Leadership-RN Basic scenario. The second and all following factors explained smaller and smaller portions of the variance and were all uncorrelated with each other.

This evidence supported that one component or latent factor, "clinical competency," exerts fundamental influence on the observed variables in Quantum. Clinical competency is latent in the sense that it is assumed to actually exist in the student's integrated KSAs and clinical reasoning ability, but cannot be measured directly. However, it does exert influence on the student's performance to the evaluation criteria that constituted the scenario's rating sheet.

These results indicate the sets of evaluation criteria within and between each subdomain can be combined into a single construct – clinical competency – consistent with the scoring structure envisioned by the SMEs. This unidimensionality evidence results in additivity [49]. Additivity refers to the properties of the measurement units. Smith (2002) pointed out these units are called logits (logarithm of odds) and have the desirable property of maintaining the same size (i.e., interval) over the entire continuum [50]. Since the combination of parameters is additive, this implies that Quantum's parameters (e.g., instructor severity, evaluation criterion difficulty, student ability, and rating scale thresholds) can be expressed as real numbers on a common interval scale.

Content Validation

Evidence based on test content featured logical and empirical analyses of the adequacy with which the test content represented the content domain and the relevance of the content domain to the proposed interpretation of test scores [10]. Content relevance, technical quality, and representativeness derived evidence in support of content validity [51].

Content Relevance

Content relevance was derived from SME's logical analysis of current references to support the test development process. The Process Model involved SMEs by core nursing course in all test development phases to ensure that each clinical scenario's content was relevant with adequate breadth and depth.

Technical Quality

Item analyses of the technical quality of the evaluation criteria was addressed via index of difficulty and index of discrimination in CTT or fit statistics in Rasch. Table 3 showed that all Infit mean square (MnSq) statistics were below 1.40 and greater than 0.65. These values of MnSq statistics were reasonably close to the value of 1, the expected value of the MnSq statistics when there is perfect fit between data and model [45], indicating that the data had good fit with the Rasch model for all clinical scenarios.

Table 3: Goodness of Fit to the Rasch Model, Eigenvalues and Reliability Estimates

	RN					**LPN/LVN**				
Course	Range of infit Mnsq	Eigenvalue of the 2nd prinicipal component	# of items	KR-20	ICC	Range of infit Mnsq	Eigenvalue of the 2nd prinicipal component	# of items	KR-20	ICC
Health Assessment										
Basic	0.66-1.38	0.9615	41	0.76	0.92					
Intermediate	0.77-1.37	0.9326	110	0.89	0.97					
Advanced	0.81-1.27	0.9185	119	0.90	0.99					
Fundmentals										
Basic	0.66-1.40	0.9452	43	0.77	0.94	0.65-1.40	0.9131	46	0.79	0.82
Intermediate	0.72-1.39	0.9295	56	0.80	0.95	0.72-1.38	0.9012	61	0.81	0.82
Advanced	0.74-1.36	0.9011	92	0.83	0.90	0.74-1.36	0.9006	95	0.83	0.87
Medical/Surgical										
Basic	0.69-1.40	0.8667	92	0.83	0.87	0.69-1.40	0.8611	95	0.84	0.91
Intermediate	0.72-1.34	0.8555	110	0.87	0.94	0.72-1.31	0.8497	113	0.87	0.94
Advanced	0.77-1.20	0.8263	126	0.89	0.83	0.82-1.21	0.8281	128	0.89	0.99
Womesn's Health										
Basic	0.63-1.34	0.9176	66	0.75	0.82	0.65-1.34	0.8828	69	0.76	0.84
Intermediate	0.68-1.29	0.8936	94	0.81	0.87	0.68-1.29	0.8601	97	0.91	0.89
Advanced	0.72-1.27	0.8740	110	0.86	0.88	0.72-1.27	0.8491	113	0.97	0.90
Pediatrics										
Basic	0.71-1.36	0.9364	56	0.75	0.87	0.71-1.36	0.9577	58	0.79	0.83
Intermediate	0.72-1.33	0.9222	71	0.77	0.88	0.72-1.33	0.9295	73	0.82	0.87
Advanced	0.82-1.28	0.9037	100	0.84	0.90	0.82-1.28	0.9144	102	0.85	0.89
Mental Health										
Basic	0.67-1.40	0.9291	66	0.81	0.85	0.66-1.40	0.9302	67	0.76	0.81
Intermediate	0.70-1.36	0.9037	89	0.86	0.84	0.70-1.34	0.9095	89	0.82	0.82
Advanced	0.72-1.34	0.8906	114	0.88	0.84	0.72-1.30	0.8926	114	0.83	0.82
Leadership										
Basic	0.70-1.36	0.9785	48	0.85	0.79					
Intermediate	0.79-1.24	0.9556	65	0.89	0.91					
Advanced	0.87-1.19	0.9321	71	0.90	0.83					

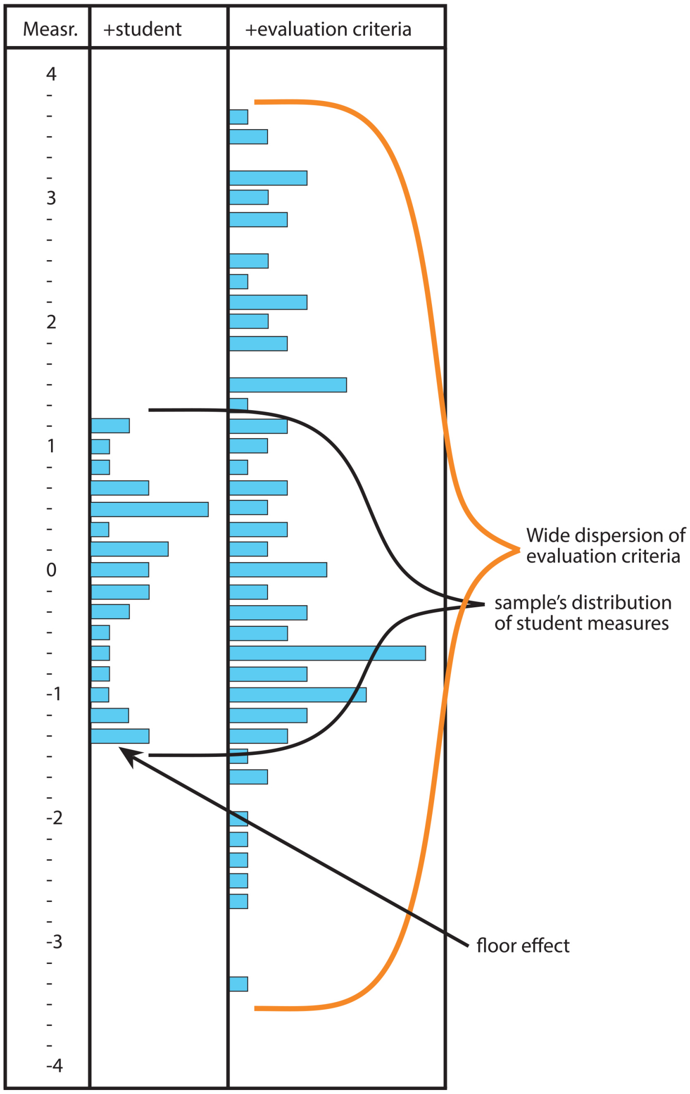
Content Representativeness

Figure 2 showed an example of good content representativeness for clinical scenario – Nurse Leadership and Management of Patient Care: a wide dispersion of evaluation criteria relative to the student measures; and no content gaps. Gaps in content representativeness were examined empirically by using the unique ordering of evaluation criterion difficulties and their individual standard errors as suggested by Smith (2001) [42]. Content representativeness analysis was repeated for each clinical scenario by course to ensure student ability distribution coverage by evaluation criteria.

Responsiveness

Figure 2 reflected two features: a floor effect occurring for this sample's ability distribution and a wide dispersion of evaluation criteria calibrations beyond the highest student measure indicating that this clinical scenario has the capacity to be responsive to an intervention. This illustrated how Quantum is equipped to compare a student's test score before and after the intervention to support the external aspect of validity [52].

Figure 2. Variable Map of Nurse Leadership and Management of Patient Care Clinical Scenario



Reliability

As recognized by the *Standards* [10], the level of reliability of test scores has implications for the validity of test score interpretations. Table 3 showed two reliability estimates for equivalence and internal consistency. An appropriate technique to express equivalence (inter-rater agreement) is an intra-class correlation (ICC) based on empirical evidence from the *Rating Scale Analysis*. Internal consistency refers to measurement error introduced by variations in evaluation criteria content and quality; the validity evidence supports the Kuder–Richardson Formula 20 (KR-20) as a measure for dichotomous items.

For all clinical scenarios, inter-rater agreement estimates were excellent according to research published by Cicchetti in 1994 with ICC estimates ranging from 0.79 to 0.99 [53]. An ICC estimate lower than 0.70 indicates the instructors have a substantial amount of disagreement among them. Continuing, internal consistency estimates using KR-20 ranged from acceptable (0.75) to excellent (0.97) according to research published by Salvucci, Walter, Conley, Fink, and Saba in 1997 [54]. Internal consistency estimates lower than 0.70 are considered questionable. The KR-20 for LPN/LVN Women's Health – Advanced clinical scenario showed very high reliability (0.95 or higher) indicating that several evaluation criteria were redundant, but deemed necessary by the SMEs.

Invariance

Analyses on student measure invariance over evaluation criteria were conducted and presented in Table 3. As the clinical scenario unfolded new evaluation criteria were added. There was improved fit between the data and Rasch model (MnSq values reasonably close to 1), which yielded the same student measures within measurement error. The results indicated that previously obtained student measures are generalizable over the new evaluation criteria.

Consequential

The application of cut scores has a direct relationship with the consequential aspect of validity [46]. Quantum features summative feedback that would include a cut score to distinguish between performance levels achieved with a course grade; and formative feedback with the intent to improve learning and future performance [15, 27].The development of competence depends upon students receiving formal feedback as part of a continuous assessment cycle [55]. Quantum provides definitive feedback and customized remediation linked to the evaluation criteria based on content validity evidence previously described to enable students to identify their strengths and weaknesses, and shows them how to improve where weak or build upon what they do best. Definitive feedback on current student performance is a crucial part of the debriefing process [56].

Discussion

Simulation and the diversity of nursing care has created a need for a standardized, performance assessment test to measure a student performance across all three domains of learning, just as standardized tests by ATI, HESI, and Kaplan are needed to measure knowledge at the end of each semester. These results provide important information about the usefulness of information provided by Quantum.

One, many researchers treat assigned ratings as interval measurement scales, but we were able to empirically validate the rating scale functionality. This validated interval measurement scale supports mathematical operations needed to calculate means and standard deviations [57, 58].

Two, Quantum's capability to estimate and adjust for instructor differences in deriving student test scores is an advantage over other performance assessment tests because the CTT method of basing decisions for rated performances directly on raw scores is unfair to students who encounter severe instructors. Meaning, CTT approaches such as percentage correct, Kappa, and G-theory do not address the issue of construct validity and offer no solution to the problem of obtaining error-adjusted test scores. As Banerji (2000) noted, the lack of rater agreement on tasks can be identified by CTT analysis but not accounted for in deriving student test scores, a factor that affects test score validity [34]. The capability to estimate and adjust for instructor differences with Quantum is thus an important advantage to objectively measuring a student's clinical competence.

Three, in order to comply with measurement best practices [10], after each test administration equivalence and internal consistency reliability estimates will be calculated to address the major sources of error variance (e.g., instructors, evaluation criteria) in Quantum that may undermine test score interpretation. Equivalence and internal consistency reliability estimates will be reported on the Administrative Report to provide nurse educators with confidence to interpret students' test scores knowing the reliability of measures.

Four, Quantum can provide diagnostic statistics to detect a variety of rater effects outlined by Gaberson, Oermann, and Shellenbarger (2015) [59]. For example, a rater bias analysis can be run to detect patterns in data that may be indicative of differential instructor functioning over time. This analysis may appeal to academic stakeholders who need to employ large numbers of clinical adjunct instructors.

Finally, based on relevant validity evidence, researchers were able to demonstrate the suitability of test scores derived from Quantum for a given decision-making context in order to provide nurse educators with a valid option beyond multiple-choice exams [2].

Limitations

Limitations of this study include collecting sufficient statistics prior to shifting from two instructors to one instructor for each student performance. In addition, instructors' suggested a toggle switch to move easily between subdomains and evaluation criteria on the in-App rating sheet and an application that functions on the latest iPad.

Implications for Nursing Education

Quantum's outcome assessment data addresses both the student success and learning gains that will result from simulation education. Instructors can observe and evaluate students "soft" benefits, such as their ability to: use nursing process as a means of decision making; and synthesize the information from the scenario with nursing theory in preventive health care.

Student Comments

Regarding learning gains and success, students liked receiving feedback on their performances in simulation. Students reported the value of this learning experience, which spoke to engagement: "…it helped me to think about what I should be looking for ahead of time;" "…it helped me bring it all together;" and "…helped me to think critically." With definitive feedback, the greater the potential for learning [60], and it is possible to pull apart key pieces during debriefing that can serve as building blocks to teach students how to think like a nurse [56].

Score reports are particularly relevant after passing NCLEX and graduating. Score reports can be used by graduate nurses in their professional portfolio to support any claims made on their resumes or in interviews to prove their past achievements. Overall, definitive feedback helps promote student learning, allows the evaluation of students and course curricula, and permits documentation of clinical competence stages [61].

Conclusion

The discipline of nursing is a hands-on profession requiring an unbiased, instructor-mediated, performance assessment test to measure a student's performance objectively. Moving beyond multiple-choice tests, Quantum addresses many challenges and issues inherent in a performance assessment measuring student competency. It provides construct validity evidence and offers a solution to obtaining error-adjusted test scores. The introduction of Quantum removes the barriers related to time and workload in an academic environment required to develop psychometrically sound, performance assessment tests. Thereby, alleviating the burden on faculty or facilitators whose responsibility it is to evaluate a student's performance, differentiate competency, and assign grades. This study supports the use of Quantum as an instructor-mediated, performance assessment test, which consistently measures learner competence and provides a holistic view of the student in either formative or summative experiences. Quantum offers faculty confidence in consistent, objective evaluation ensuring that graduates are safe and competent.

References

1. Miller GE (1990) The assessment of clinical skills/ competence/performance. Academic Medicine 65: S63-S67.
2. Benner P, Sutphen M, Leonard V, Day L (2010) Educating nurses: A call for radical transformation. San Francisco: Jossey Bass.
3. Hayden JK, Smiley RA, Alexander M, Kardong-Edgren S, Jefferies PR (2014) The NSCBN National Simulation Study: A longitudinal, randomized, controlled study replacing clinical hours with simulation in prelicensure nursing education. Journal of Nursing Regulation 5.
4. Leigh G, Stueben F, Harrington D, Hetherman S (2016) Making the case for simulation–based assessments to overcome the challenges in evaluating clinical competency. International Journal of Nursing Education and Scholarship 13: 1-8.
5. Kardong-Edgren S, Adamson KA, Fitzgerald C (2010) A review of currently published evaluation instruments for human patient simulation. Clinical Simulation in Nursing 6: e25-e35.
6. Collins S, Callanhan MF (2014) A call for change: Clinical evaluation of student registered nurse anesthetists. AANA Journal 82: 65-72.
7. Solnick A, Weiss S (2007) High fidelity simulation in nursing education: A review of the literature. Clinical Simulation in Nursing Education 3: e41-e42.

8. Walsh M, Bailey PH, Koren I (2009) Objective structured clinical evaluation of clinical competence: An integrative review. Journal of Advanced Nursing 65:1584-1595.
9. Stewart BJ, Archibold PG (1997) A new look for measurement validity. Journal of Nursing Education 45: 204-211.
10. American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], & Joint Committee on Standards for Educational and Psychological Testing. (2014). Standards for educational and psychological testing. Washington, DC: AERA.
11. Wolfe EM, Smith EV Jr. (2007a) Rasch instrument development and measure validation. Journal of Applied Measurement 8: 97-123.
12. Li S (2007) The role of simulation in nursing education: A regulatory perspective.
13. Brink Y, Louw QA (2012) Clinical instruments: reliability and validity critical appraisal. Journal of Evaluation in Clinical Practice 18: 1126–1132.
14. Jefferies P (2007) Simulation in nursing education. New York: National League for Nursing.
15. Lioce L, Meakim CH, Fey MK, Chmil JV, Mariani B, et al. (2015) Standards of best practice: Simulation Standard IX: simulation design. Clinical Simulation in Nursing 11: 309-315.
16. Toth J (2011) Development of the Basic Knowledge Assessment Tool for Medical-Surgical Nursing (MED-SURG BKAT)© and implications for in-service educators and managers. Nursing Forum 46: 110-116.
17. Meakim C, Boese T, Decker S, Franklin AE, Gloe D, et al. (2013) Standards of best practice: Simulation Standard I: Terminology. Clinical Simulation in Nursing 9: s3-s11.
18. Ericsson K, Simon H (1980) Verbal reports as data. Psychological Review 87: 215-251.
19. Slameca NJ, Fevreiski J (1983) The generation effect when generation fails. Journal of Verbal Learning and Verbal Behavior 22: 153-163.
20. Masters G, Forster M (1996) Progress Maps. (Part of the Assessment Resource Kit). Melbourne, Australia: The Australian Council for Educational Research, Ltd 1-58.
21. Flowers C, Browder D, Wakeman S, Karvonen M (2007) Links for academic learning: The conceptual framework. National Alternative Assessment Center (NAAC) and the University of North Carolina at Charolette.
22. Dewey J (1933) How we think: A restatement of the relationship of reflective thinking to the educative process (2nd rev. ed.). Lexington, MA: D. C. Heath.
23. Sergiovanni TJ (1986) Understanding reflective practice. Journal of Curriculum and Supervision 1: 353-359.
24. Seifert KL (1999) Reflective thinking and professional development. Boston: Houghton-Mifflin Company.
25. Centers for Disease Control. (2002) Guideline for hand hygiene in health care settings.
26. Quality and Safety Education for Nurses Institute. (n.d.). Pre-licensure KSAs. Retrieved on April 16, 2011, from http://qsen.org/competencies/pre-licensure-ksas/
27. Sando CR, Coggins RM, Meakim C, Franklin AE, Gloe D, et al. (2013) Standards of Best Practice: Simulation Standard VII: Participant Assessment and Evaluation. Clinical Simulation in Nursing 9: S30-S32.
28. American Association of Colleges of Nursing. (2008) The Essentials of Baccalaureate Education for Professional Nursing Practice.
29. National League for Nursing. (2012) Outcomes and Competencies for Graduates of Practical/Vocational, Diploma, Baccalaureate, Master's Practice Doctorate, and Research Doctorate Programs in Nursing. (1st ed.). New York: Author.
30. National Council of the State Board of Nursing (2016) 2016 NCLEX-RN Test Plan.
31. Rossomando v. Board of Regents of the University of Nebraska, 2 F.Supp.2d 1223 (1998).
32. Grbach W Reformulating SBAR to "I-SBAR-R." (n.d.).
33. Chatterji M (2003) Designing and using tools for educational assessment. Boston, MA: Allyn & Bacon.
34. Banerji M (2000) Construct validity of scores/measures from a developmental assessment in mathematics using classical and many-facet Rasch measurement. Journal of Applied Measurement 1: 177-198.
35. Hetherman SC (2004) An application of multi-faceted Rasch measurement to monitor effectiveness of the Written Composition Test of English in the New York City Department of Education. New York: Teachers College, Columbia University. Published Dissertation.
36. Crocker L, Algina J (1986) Introduction to classical and modern test theory. Orlando, FL: Harcourt Brace Jovanovich.
37. Rasch G (1960) Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Danish Institute for Educational Research. (Expanded edition, 1980. Chicago: University of Chicago Press.)
38. Linacre JM (1994) Many-facet Rasch measurement. Chicago: MESA Press.
39. SPSS Inc. Released (2007) SPSS for Windows (Version 16.0) [Computer software]. Chicago, SPSS Inc.
40. Linacre JM (2005) Winsteps® (Version 3.55) [Computer software]. Chicago: Winsteps.com
41. Linacre JM (2003) FACETS-Rasch measurement computer program (Version 3.47) [Computer program]. Chicago: Winsteps.com
42. Smith EV Jr (2001) Evidence of the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. Journal of Applied Measurement 2: 281-311.
43. Linacre JM (2004) Optimal rating scale category effectiveness. In E.V. Smith, Jr. and R.M. Smith (Eds.), Introduction to Rasch measurement (pp.258-278). Maple Grove, MN: JAM Press.
44. Linacre JM (2002) Optimizing rating scale category effectiveness. Journal of Applied Measurement 3: 85-106.
45. Linacre JM (2014b) Winsteps® Rasch measurement computer program user's guide. Beaverton, OR: Winsteps.com.
46. Wolfe EM, Smith EV Jr. (2007b) Rasch instrument development and measure validation. Journal of Applied Measurement 8: 204-234.
47. Cattell RB (1966) The scree test for the number of factors. Multivariate Behavioral Research 1: 629-637.
48. Kaiser H F (1960) The application of electronic computers to factor analysis. Educational and Psychological Measurement 20: 141-151.
49. Linacre JM (1998) Detecting multidimensionality: Which residual data-type works best? Journal of Outcome Measurement 2: 266-283.
50. Smith EV Jr. (2002) Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. Journal of Applied

Measurement 3: 205-231.

51. Messick S (1989) Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp.13-103). New York: Macmillan.
52. Medical Outcomes Trust Scientific Advisory Committee (1995) Instrument review criteria. Medical Outcomes Trust Bulletin 1-4.
53. Cicchetti DV (1994) Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychological Assessment 6: 284-290.
54. Salvucci S, Walter E, Conley V, Fink S, Saba M (1997) Measurement error studies at the National Center for Education Statistics (NCES). Washington D. C.: U. S. Department of Education.
55. Heaslip V, Scammell JME (2012) Failing underperforming students: The role of grading in practice assessment. Nurse Education in Practice 12: 95-100.
56. Decker S, Fey M, Sideras S, Caballero S, Rockstraw L, et al. (2013) Standards of best practice: Simulation standard VI: The debriefing process. Clinical Simulation in Nursing 9: S26-S29.
57. Merbitz C, Morris J, Grip JC (1989) Ordinal scales and foundations of misinference. Archives of Physical Medicine and Rehabilitation 70: 308-312.
58. Wright BD, Linacre JM (1989) Observations are always ordinal: measurements, however, must be interval. Archivers of Physical Medicine and Rehabilitation 70: 857-860.
59. Gaberson KB, Oermann MH, Shellenbarger T (2015) Clinical teaching strategies in nursing. New York, NY: Springer.
60. Neary M (2000) Responsive assessment of clinical competence: part 2. Nursing Standard 22: 35-40.
61. Benner P (1984) From novice to expert: Excellence and power in clinical nursing practice. Menlo Park: Addison-Wesley 13-34.