ANALYSIS OF RATER RELIABILITY USING A FACULTY DEVELOPED, REVISED OSCE EVALUATION INSTRUMENT

by

TAMARA ANDREWS

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON
August 2016

Acknowledgements

I would like to express my sincere gratitude to my dissertation committee. Thank you enormously to Dr. Judy LeFlore for the leadership, guidance, mentoring and support that you provided over the past two years as my Dissertation Chair. I look to you as a role model of character and excellence, and appreciate the time spent on my behalf. To Dr. Jennifer Gray, committee member, I thank you for your guidance, your knowledge and your devotion to seeing me succeed, and for your attention to detail with the final manuscript and presentation. Thank you Dr. Patricia Thomas, committee member as you were always available to help me locate information or a reference, and gave continuing encouragement. I appreciate all of your help.

I would like to thank my family who supported and prayed for me: my brother Bobby Andrews and sister-in-law Lydia Andrews. Thank you for believing in me. My mother Avonelle Andrews and father Robert Andrews were the inspiration for me to persevere to the finish line. My mother was so excited for me to succeed with the Ph.D. I wish you could be here to see that I have fulfilled this dream, but know you are watching from above.

I would also like to thank Dr. Denise Cauble for your friendship, collegiality and support during our mutual journey toward achieving this goal. Your encouragement and support helped me continue when I was discouraged. I will never forget our Saturday sessions at UTA with Starbucks coffee and snacks. I appreciate Dr. Susan Baxley, who mentored and motivated me to succeed. You have gone the extra mile to be my chief supporter and assure my success. You continued to motivate me to keep my eye on the prize. Thanks to the members of Cohort 1 of the UTA PhD program for supporting me and celebrating my success.

I could not have been successful without the continued assistance and guidance of Vivian Lail-Davis, Administrative Assistant. Thank you for keeping me on track and providing reminders so I did not miss a deadline, and encouraging me to succeed.

Thank you to all my colleagues from the University of Texas at Arlington College of Nursing and Health Innovation. You encouraged me, cheered me on, and celebrated my success, I will never forget you and thank you for the role you played in helping me achieve my goal.

August 5, 2016

Abstract

ANALYSIS OF RATER RELIABILITY USING A FACULTY DEVELOPED, REVISED OSCE EVALUATION INSTRUMENT

Tamara Andrews, PhD

The University of Texas at Arlington, 2016

Supervising Professor: Judy LeFlore

Nurse educators need valid and reliable instruments to evaluate student performance during summative clinical experiences and Objective Structured Clinical Examinations (OSCEs). The purpose of this retrospective, secondary data analysis was to analyze existing data to determine whether there were differences between raters using a faculty-developed, revised OSCE instrument designed to evaluate student performance in a summative course OSCE. Data were extracted from an existing database collected during a prior study to compare OSCE instruments, within a large, urban college of nursing. Data were examined for a sample of 44 subjects whose OSCE performances were rated by faculty using the revised OSCE instrument. Differences between course faculty and non-course faculty were examined for four categories of the OSCE instrument including patient safety, assessment, planning and medication administration, to determine whether the revised OSCE instrument demonstrated inter-rater reliability. In this study, there were significant differences between rater groups in ratings of the subject's performances. There was no agreement between rater groups. Probable reasons for these differences were explored, including grade inflation, failure to fail and rater error. Nurse educators need to determine whether to continue to use faculty developed instruments or to select evaluation instruments which have been tested and shown to be reliable and valid.

Table of Contents

List of Illustrations

List of Tables

Chapter 1

Introduction

Nurse educators are challenged to develop learning activities that facilitate learning and acquisition of skills. During prelicensure education, nursing students need to demonstrate competency in performing a variety of technical skills, and the ability to make sound clinical judgments in caring for patients. Simulation and objective structured clinical examinations (OSCEs) are used in undergraduate nursing curricula as both teaching and evaluation strategies.

Simulation is described as activities designed to mimic reality in a clinical environment and allows students to practice and demonstrate competency in performing technical skills, develop clinical decision-making skills, and learn to recognize and manage a change or deterioration in a patient's clinical status (Raurell-Torreda, Olivet-Pujol, Romero-Collado, Malagon-Aguilera, Patino-Maso, & Baltasar-Bague, 2015; Cant & Cooper, 2010; Doolen, 2015; Cooper, Cant, et al., 2015; Merriman, Stayt, and Ricketts, 2014; Fisher & King, 2013; Gordon & Buckley, 2009). Simulation includes the use of high-fidelity simulation manikins, standardized patients, task trainers, and scenarios designed by nursing faculty that allow students to encounter situations seldom found in clinical practice and to practice in a safe environment without risk of harm to a patient (Mikasa, Cicero & Adamson, 2013; Raurell-Torreda, et al., 2015).

An objective structured clinical examination (OSCE) is a performance-based examination where students are observed demonstrating specific clinical behaviors to assess their ability to apply theoretical knowledge to a clinical situation (Harden, Stevenson, Downie, & Wilson, 1975; Jones, Pegram & Fordham-Clarke, 2010; & McWilliam & Botwinski, 2010). OSCEs have been used in formative evaluation, to improve learning, and provide feedback on strengths and weakness and in summative evaluation, to objectively measure student performance, as well as evaluate course and program outcomes (Cazzell & Howe, 2012; Rentschler, Eaton, Cappiello, McNally, and McWilliam, 2007; and Traynor & Galanouli, 2015). OSCEs are often designed to be implemented using simulation.

When using OSCEs and/or simulation to evaluate student performance, it is crucial to establish reliability and validity of the rating tool to ensure that the tool measures the behavior it is intended to measure, for the specific population involved (Cazzell & Howe, 2012; Kardong-Edgren, Adamson & Fitzgerald, 2010). Evaluation of nursing performance must include the cognitive, psychomotor and affective domains (Cazzell & Howe, 2012; Kardong-Edgren, Adamson & Fitzgerald, 2010). Therefore tools designed to evaluate student performance should be tested and validated to measure each of these domains (Hayden, Keegan, Kardong-Edgren & Smiley, 2014a).

Background and Significance of the Problem

Research has indicated that registered nurses' ability to recognize changes in a patient's condition is crucial to positive patient outcomes. Failure to recognize these changes and act quickly, puts patients at risk for poor outcomes. (Beaumont, Luettel, & Thomson, 2008; Minick & Harvey, 2003; Purling & King, 2012; Shever, 2011). Much has been written about the concept of failure to rescue, which has also been used to describe nurses' failure to recognize and respond in a timely manner to a change in clinical condition (Aiken, Clarke, Cheung, Sloan, & Silber, 2003; Aiken, Clarke, Sloane, Sochalski, & Silber, 2002; Clarke & Aiken, 2003 & Schubert, 2012). Adding to the patient safety concerns are the fact that new graduate nurses may not yet have developed clinical reasoning skills, including the ability to recognize signs of clinical deterioration (Cooper, Kinsman, Buykx, McConnell-Henry, Endacott, & Scholes, 2010; Endacott, Scholes, Buykx, Cooper, Kinsman, & McConnell-Henry, 2010). Benner, Sutphen, Leonard & Day (2010) documented a gap in the ability of new graduates to apply theoretical knowledge to practice.

The National Council of State Boards of Nursing (NCSBN) conducted a three-year project (Hayden, Keegan, Kardong-Edgren, & Smiley, 2014a; Hayden, Smiley, Alexander, Kardong-Edgren & Jeffries, 2014b) exploring use of simulation for high stakes evaluation in pre-licensure RN programs. In the National Simulation Study, the NCSBN defined clinical competency as

> the ability to observe and gather information, recognize deviations from expected patterns, prioritize data, make sense of data, maintain a professional response demeanor, provide clear communication, execute effective interventions, perform

> nursing skills correctly, evaluate nursing interventions, and self-reflect for performance improvement within a culture of safety (Hayden, et al, 2014a, p. S42).

The NCSBN supports use of simulation in nursing education, based on evidence that high-quality simulation experiences can be substituted for up to half of traditional clinical hours, and can be used to evaluate clinical competency during high stakes evaluation (Hayden, et al., 2014a; Hayden, et al., 2014b; Rizzolo, Oermann, Jeffries & Kardong-Edgren, 2011).

OSCEs are often employed by nursing faculty to guide clinical experiences in the final clinical capstone course, and to evaluate nursing students' ability to apply theoretical knowledge to actual, clinical or simulated experiences (McWilliam & Botwinski, 2010). Additionally OSCEs are used as a summative program evaluation, which ultimately determines whether the student will progress to graduation (Jeffries & Norton, 2005; Mitchell, Henderson, Groves, Dalton & Nulty, 2009; Rentschler, et al., 2007; & Rushforth, 2007). An OSCE using clinical simulation allows educators to use the same scenario under controlled conditions, to evaluate students in a consistent manner (Kardong-Edgren, Adamson & Fitzgerald, 2010).

Nurse educators need reliable and valid tools to evaluate students' clinical competency and performance during high-stakes testing involving simulation and OSCEs, but there is a lack of such tools that have undergone rigorous testing for reliability and validity (Hayden, et al., 2014a; Kardong-Edgren, Adamson & Fitzgerald; 2010; Cordi, Leighton, Ryan-Wenger, Doyle & Ravert, 2012). One reason to have valid and reliable tools is to minimize the likelihood of rater error. Rater error can be categorized as leniency, inconsistency, the halo effect, and restriction of range (Iramaneerat & Yudkowsky, 2016). Leniency is defined as the tendency for raters to give subjects higher ratings than what they have earned. At the other end of this spectrum are the severe raters who assign consistently lower ratings. For both lenient and severe raters, the term leniency error can produce inaccurate and unfair results. Inconsistency refers to the rater's tendency to assign ratings with more randomness, indicating a lack of understanding of the rating criteria. The halo effect is when raters allow an individual subject's performance in one trait to influence the evaluation of performance in other traits. This is believed to occur when a rater's

general impression of a subject influences the ratings of performance in other areas. Finally, the restriction of range error occurs when raters tend to arbitrarily assign ratings to one end of the scale (lenient or severe), or to the midpoint of the scale, failing to use ratings in other areas of the scale (Iramaneerat & Yudkowsky, 2016).

This retrospective, secondary analysis was conducted to determine whether a newly developed OSCE rating tool demonstrated inter-rater reliability when used by multiple raters.

Development of the Framework

Tanner Model Overview

The framework for this study is based upon the Tanner Clinical Judgment Model (Tanner, 2006) which depicts the processes a student would undertake during a patient clinical interaction. These processes reflect the dimensions the revised OSCE instrument was designed to evaluate. Tanner developed her model based upon research on clinical judgment and clinical decision-making. Included in that research were conclusions from Patricia Benner's qualitative studies of nurses to explore their experiences with clinical judgment through interpretation of narrative accounts (Benner, et al., 2010). Tanner's model begins with contextual concepts based on what the nurse brings to the encounter. Those are described as the context, background, and relationship to the patient. The background includes the nurse's level of knowledge and experience. Experienced nurses have encountered patient situations and have learned to recognize patterns that direct their clinical judgment; however, beginning nurses must develop these analytical reasoning skills over time (Cioffi, 2000). A fundamental concept to that model was the specific relationship that a nurse had with the patient, or the extent to which the nurse "knows" the patient. This knowing is based on interactions, communication, learning about their lives, and developing a rapport and insight into their clinical situation (Tanner, Benner, Chesla & Gordon, 1993). Finally, the context includes the specific situation in which the nurse is caring for the patient, including the nursing unit, unit culture, personnel, and specific patient details that help nurses prioritize and make decisions.

Tanner's complete model includes the concepts of Noticing, Interpreting, Responding and Reflecting (Tanner, 2006). Noticing includes the elements of context, background and relationship, expectations and initial grasp. The concept of Interpreting includes the elements of reasoning patterns, analytic, intuitive, and narrative. The concept of Responding includes the elements of action and outcomes. The concept of Reflecting includes reflection-in-action that occurs during the interpreting phase, followed by reflection-on-action and clinical learning.

The concepts in Tanner's model most relevant to this study include: 1) Noticing; 2) Interpreting; and 3) Responding. The concept of Reflecting was not included as part of this study. In this study, the Framework used the concept of Noticing to include the actual physical assessment of the patient, but additionally included the nurse's ability to use this "initial grasp" of the situation (Figure 1). Initially a student nurse would encounter a patient, whether in an actual clinical setting or in a simulated scenario, and be given a shift report of the patient's presenting condition. In order to formulate a plan of care for the patient, the student nurse begins with an assessment, typically a focused assessment, specific to the patient's problem. A focused assessment is defined as one that is smaller in scope than a complete or comprehensive assessment, often centered on one body system or symptom (Jarvis, 2012). The focused assessment consists of both subjective and objective information. The student nurse conducts a brief interview of the patient, to explore their chief complaint, ascertain the patient's ability to participate in the interview, and some basic information about their level of consciousness and mental status. Then the student nurse performs a focused physical assessment specific to the patient's presenting condition, and including major body systems (Jarvis, 2012).

As the nurse comes to know a patient, Noticing includes the patient's typical responses to their disease process and variations from the norm (Tanner, 2006). Interpreting and responding describes the nurse's ability to immediately synthesize this data, before forming initial and final hypotheses until the appropriate course of action is determined. Interpreting and Responding constitute the nurse's clinical judgment or clinical reasoning. Based on the subjective and objective data, the student nurse recognizes either that there is a need for action, or that no

immediate action is needed. If action is needed, the student nurse is expected to correctly respond to the situation in order that the patient's condition improves or stabilizes.

In Tanner's model, the last concept is Reflecting. In this current study, the Framework varies slightly in that the nurse must go back to Noticing, which entails a reassessment of the patient's condition and the evaluation of effectiveness of interventions.

Framework for Current Study

This Framework was developed to capture a specific piece of the Tanner model, that of performing a focused assessment, promptly recognizing patient problems; and responding appropriately. Because of the importance of ongoing assessment and reassessment, the concept of Noticing was included at the end of the sequence, in the context of reassessment to evaluate effectiveness of actions or need to implement different interventions. See Figure 1-1 below.

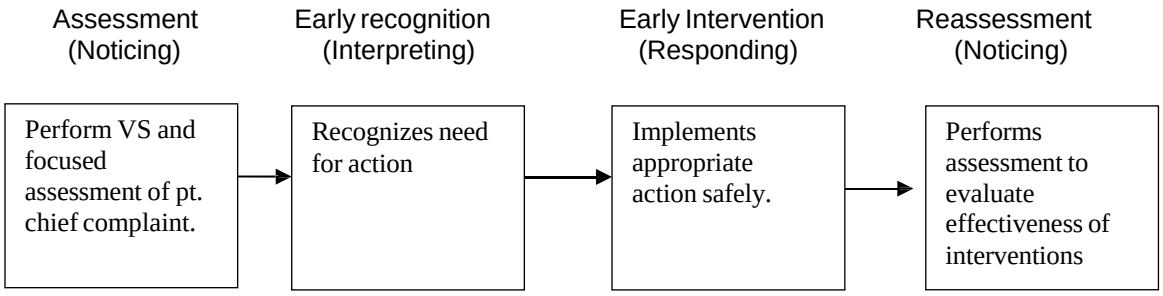

Figure 1-1 Study Framework

Because early recognition is a vital aspect of nursing care, it must be taught and evaluated during prelicensure nursing education, as well as continuing to be reinforced as the nurse enters the professional role. Minick & Harvey (2003) found that nurses often did not learn these skills during their educational preparation, but later, during their professional practice. Recognizing and responding to a patient's changing condition is an essential skill, and is well suited for practice during clinical simulation in a controlled environment (Gordon & Buckley, 2009; Endacott, et al., 2010; Cooper, et al., 2010). Many assessment tools have been developed to evaluate performance in simulation, but few have undergone rigorous testing for validity and reliability (Hayden, et al., 2014a; Hayden, et al., 2014b). Typically simulation is performed by students working in groups and the simulation tools are used to evaluate students working in

teams (Hayden, et al., 2014a; Adamson, Kardong-Edgren, & Willhaus, 2013; & Cordi, et al., 2012). OSCEs are more commonly used to evaluate individual clinical performance as part of a summative evaluation. Nurse faculty members need reliable and valid OSCE instruments. The OSCE instrument that is the focus of this study will be used to capture the elements of this framework. A summary of definitions of key concepts of the study are provided in Table 1-1.

Table 1-1 Conceptual Definitions

Concept or Process	Definition for this Study
Assessment	The systematic and ongoing collection of patient data, performed to develop, implement and revise the patient's plan of care. American Nurses Association, 2015).
Focused assessment	An assessment that is narrower in scope. It consists of brief, specific subjective data, and physical examination of areas necessary to explore the patient's presenting condition. (Jarvis, 2012).
Early recognition/detection of patient problems	Identifying and acting on subtle, difficult-to-describe patient changes that signal a potential patient problem. (Minick & Harvey, 2003; Cooper, Cant, Porter, Missen, Sparkes, McConnell-Henry, & Endacott, 2013; Fisher & King, 2013).
Deterioration of patient condition	Detectable changes in physiological signs and symptoms (Rich, 1999).
Clinical judgment	An interpretation or conclusion about a patient's needs or concerns and the decision of whether or not to take actions. (Tanner, 2006).
Simulation	Activities that closely represent the reality of a clinical environment; designed to allow demonstration of skills, decision-making and clinical reasoning. (Raurell-Torreda, et al, 2015; Cant & Cooper, 2010).
OSCE	A performance-based examination where students are observed demonstrating various clinical behaviors in order to assess transfer of theoretical knowledge to practice. (McWilliam & Botwinski, 2010).

Study Purpose

The purpose of this retrospective, secondary data analysis is to analyze existing data to determine whether there are differences between raters using a faculty-developed, revised OSCE instrument designed to evaluate student performance in a summative course OSCE. Based on the study framework, faculty, using the OSCE instrument, need to be able to ascertain whether the student nurse met expectations to perform a focused assessment, recognize the patient problem, respond immediately to assure a safe patient outcome and reassess effectiveness of

interventions The revised OSCE instrument included observable and measurable behaviors reflecting each concept and process in the study framework. The summative OSCE was designed to evaluate whether students could integrate the knowledge and skills gained throughout the program, to assess a patient, recognize changes in condition, and plan and safely care for a patient (Cazzell & Howe, 2012; McWilliam & Botwinski, 2010).

In a prior study, faculty revised three scenario-specific OSCE instruments used to evaluate the elements of assessment, reassessment, medication administration, and safety principles using simulated clinical vignettes (Stiller, et al., 2015). Data were collected using both the original and revised OSCE instruments.

This study will consist of analysis of the data from the database to determine whether the revised OSCE instrument demonstrated reliability when used by multiple faculty raters. Six faculty members from the clinical course observed student OSCE performances and used the revised OSCE instrument to rate student performance. Three Non-course Faculty members who were not assigned to the clinical course viewed video-recordings of the same students performing the OSCE, and used the revised OSCE instrument to rate student performance.

Research Questions

1. Were there differences among the six Course Faculty's ratings of students using the revised OSCE instrument?
2. Were there differences among the three Non-course Faculty's ratings of students using the revised OSCE Instrument?
3. Were there differences between Course Faculty and Non-course Faculty's ratings of students using the revised OSCE instrument?
4. What were the inter-rater reliability coefficients for each subscale on the revised OSCE instrument when scored by multiple raters?

Assumptions

1. Nursing faculty will demonstrate expertise in use of OSCE rating tools to evaluate student performance in a summative OSCE.

2. External faculty raters will be unbiased in their ratings of student performance during a summative course OSCE.

Summary

In summary, simulation is an essential tool for use in undergraduate nursing education as a teaching and evaluation strategy. Simulation presents ideal learning opportunities to assist students and educators to incorporate these both technical skills and clinical reasoning skills in a safe environment. Using well-designed summative OSCEs, facilitated by using simulation, nurse educators are able to evaluate student performance in a controlled setting, where nursing students can safely perform and learn from experience. This study will contribute to the body of nursing knowledge by providing evidence related to the rater reliability of an OSCE rating tool used to evaluate nursing students' assessment and clinical decision making skills. Nurse educators need tools to evaluate student performance during simulation and OSCEs that are valid and proven to be reliable. This study may add to the body of nursing knowledge regarding simulation, OSCEs, and establishment of the reliability of faculty-developed OSCE rating tools.

Chapter 2

Review of the Literature

Introduction

The review of literature for this retrospective, secondary data analysis was conducted using established, databases accessed through the University of Texas at Arlington online Library website. Databases included The Cumulative Index of Nursing and Allied Health Literature (CINAHL), Academic Search Complete, Health Source: Nursing/Academic Edition, Science Direct, MEDLINE, PsycArticles, and PsycINFO.

Nursing faculty need reliable and valid tools in order to evaluate student performance in key competencies such as performing a focused assessment and being able to recognize a change or deterioration in patient condition. The key words used for the search included nursing assessment; surveillance; undergraduate nursing students; recognizing deterioration of patient condition; early recognition; clinical judgment, simulation, objective structured clinical examination (OSCE); OSCE checklists; OSCE tools; inter-rater reliability; validity; evaluation of clinical performance; clinical failure; failure to fail; grade inflation; deliberate practice, and pattern recognition. The Boolean operators AND was used to combine terms, including undergraduate nursing students AND OSCE, as well as simulation AND OSCE. Inclusion criteria for the review of the literature was publications from 2000 to current, peer reviewed journals, and English language. Exclusion criteria were studies involving licensed nurses instead of nursing students, or other health disciplines aside from nursing. In addition to formal searches for current literature, additional articles were obtained through review of references from the articles retrieved during the literature search.

Review of Relevant Literature

The review of literature begins with evidence that establishes the importance of assessment and surveillance. Assessment and surveillance are two essential skills for nursing students to develop. The next section of the review is comprised of the research related to using

simulation to teach these skills and evaluate student achievement with OSCEs. The last section includes the evidence available on the psychometrics of OSCE tools.

Key Elements of Clinical Nursing Education

Nursing assessment

According to guidelines and standards of professional nursing organizations, nursing assessment is an expectation of the registered nurse. Nursing assessment and synthesis of data are essential skills common to all nurses, as recognized by nursing organizations including the American Nurses Association (2015), the American Association of Colleges of Nursing (2008), and the Texas Board of Nursing (2011). Assessment and synthesis are key processes in providing safe patient care. The Scope and Standards of Practice (ANA, 2015) lists assessment as the first step of the nursing process. When performing the assessment, the registered nurse collects comprehensive data in a systematic and ongoing process. The nurse uses ongoing assessment data to develop, implement, and revise the patient's plan of care as needed (ANA, 2015; Jarvis, 2012).

The Texas Board of Nursing (BON) Position Statement 15.28, RN Scope of Practice (2011), describes the nurse's role in assessment. Using the Position Statement, the BON developed curricular guidelines for nursing education programs to prepare competent nurses to practice safely and compassionately. The Differentiated Essential Competencies of Graduates of Texas Nursing Programs (Texas Board of Nursing, 2011), includes 25 competencies, placed in one of four nursing roles. The nursing role titled "provider of patient-centered care" includes the mandate of the nurse to determine the needs of patients, including physical, mental, cultural, and social, based upon interpretation of findings from the comprehensive nursing assessment.

Nursing surveillance

Nursing surveillance is similar to nursing assessment, but is the term more commonly found in the literature. Nursing surveillance has been defined as "the application of behaviors and cognitive processes in the systematic collection of information used to make judgments and predictions about a person's health status" (Dougherty, 1999, p. 524); as "the purposeful and

ongoing acquisition, interpretation, and synthesis of patient data for clinical decision-making" (McCloskey & Bulecheck, 1996, p. 632); and as "a process through which nurses monitor, evaluate, and act upon emerging indicators of a patient's change in status" (Kutney-Lee, Lake, & Aiken, 2009, p. 218).

Nursing surveillance has been conceptually described to include collection of data, patient assessment, recognizing, analyzing and responding to cues to make patient care decisions (Henneman, Gawlinski & Giuliano, 2012; Kelly & Vincent, 2011); continuous or intermittent monitoring of physiologic status (Kutney-Lee, Lake & Aiken, 2009); and recognizing subtle changes in a patient's condition, and taking necessary interventions (Mark & Harless, 2010).

Research studies exploring nursing surveillance have typically included retrospective reviews of hospital electronic medical records (Dochterman, et al., 2005; Fasolino & Verdin, 2015; Kutney-Lee, Lake, & Aiken, 2009; Schoneman, 2002; & Voepel-Lewis, Pechlavanidis, Burke & Talsma, 2013). Results from these studies suggested nursing surveillance was an important nursing intervention.

Nursing surveillance has been linked to detecting adverse events among hospitalized adults and children. Dochterman and Bulecheck (2004) used a hospital's electronic database to document planned nursing interventions performed by nurses, by type and frequency. Nursing surveillance as an intervention was documented in 49-50% in patients with heart failure (n=1,435 hospitalizations); hip fracture procedures (n=569 hospitalizations), and fall prevention (n=11,756 hospitalizations). Shever (2011) used data from this study to examine the relationship between nursing surveillance and failure to rescue (n=10,004 hospitalizations). She used matched pairs to compare patients with low surveillance (< 12 per day) to patients with high surveillance (> 12 per day). An increased number of failure to rescue events (n=135) were found among patients receiving less surveillance than in the group receiving more surveillance (n=31).

In a study examining the relationship between staffing levels, surveillance and pediatric postoperative serious adverse events (n=228), researchers concluded that increased surveillance, based on recognition of deterioration, likely facilitated the rescue of children in the setting, even during times of lower staffing levels (Voepel-Lewis, et al., 2013).

Recognition and early recognition of deterioration of patient condition

Early recognition is a vital component of nursing care, yet changes in a patient's condition that signal deterioration were often not detected (Beaumont, Luettell & Thomson, 2008; Cooper et al., 2010; Fero et al., 2010). In the Fero et al., study (2010), 75% of nursing students did not successfully achieve expected performance levels, and had the greatest difficulty recognizing patient problems and reporting to the physician (n=36). Cooper et al., (2010) used a mixed methods design where students participated in two simulated scenarios involving deteriorating patients (n=51). Knowledge and skills performance were assessed, and results showed that students' skills performance declined significantly in both scenarios as the patient's condition deteriorated.

A systematic review of studies of nurse recognition of patient deterioration (Douw, et al, 2015) yielded five quantitative studies and nine qualitative studies which matched their search criteria, which included the terms of recognition of deteriorating patients and "nurses' worry or concern". Most of the studies focused on nurses' involvement in rapid response teams (RRT), also called medical emergency teams (MET). In the analysis of themes found in the qualitative studies, Douw, et al., (2015) identified ten signs and symptoms corresponding to the nurses' worry or concern. These included changes in vital signs; change in mental status; agitation; pain; a term called "unexpected trajectory"; a subjective statement by the patient of feeling unwell; a subjective observation by the nurse that something did not seem quite right; and finally, a general category of "knowing without a rationale". Nurses have been shown to recognize patient deterioration through intuition rather than routine assessment in many cases (Odell, Victor & Oliver, 2006; Benner, Tanner, & Chesla, 2009). One recommendation from this systematic review

was to include the phenomenon of worry and concern to the nurse's algorithm of deciding whether to activate rapid response teams.

The one qualitative study which specifically addressed the general concept of early recognition was by Minick & Harvey (2003). Minick and Harvey (2003) conducted focus groups of medical-surgical nurses to discuss the phenomenon of early recognition. Nurses described specific instances of early recognition in the context of knowing the patient directly, through the family, or knowing something was not as expected. Nurses learned the skills of early recognition by repeatedly caring for patients with similar patterns of conditions and developing expertise with time and experience, known as pattern recognition (Benner, 1984; Dreyfus & Dreyfus; 1980; Dreyfus & Dreyfus, 1986; Ericsson, Whyte & Ward, 2007). Skill development typically occurred during actual clinical practice as a registered nurse, rather than during the basic undergraduate educational preparation. When nurses encounter particular kinds of patient situations in which they have experience, they improve their performance in similar situations and develop skills to interpret complex clinical situations (Christensen, 2010; Ericcson, Whyte & Ward, 2007). Often experienced nurses provided mentorship and guidance to assist nurses to develop these skills. Because of recent nursing shortages, fewer experienced nurses are available at the bedside, which may jeopardize the learning experience vital to growth of novice nurses (Minick & Harvey 2003).

Few U.S. studies have been conducted on nurses' ability to recognize deterioration of patients. Much of this research has been done in England. In 2007 the National Patient Safety Agency (NPSA) (2007a and 2007b) in England explored the issue of deaths submitted to the National Reporting and Learning System (NRLS) that occurred over a one-year period, and identified that 11% resulted from failure to detect or act upon clinical deterioration (n=66). Beaumont, Luettell & Thomson (2008) reported these findings and those of the National Institute for Health and Clinical Excellence (NICE) (NPSA 2007a) regarding the same issue. The NPSA formed a multi-agency group to coordinate several national and international work programs to improve patient safety. Beaumont and colleagues' (2008) review of agency findings states these

deaths were attributed to failure to monitor changes in vital signs or patient condition, failure to recognize the significance of the change in condition, and delay in obtaining medical attention. Subsequently, the NPSA (2007b) commissioned a program to identify underlying causal and contributory factors through interviews with clinicians, root cause analyses, and focus groups. The National Patient Safety Agency (2007a) developed and implemented a clinical guideline for the care of patients in acute care settings that included physiological observations, clear written monitoring plans, physiological track and trigger systems, provision for staff competencies, and a graded response strategy for patients identified to be at risk for clinical deterioration.

In a multi-center, mixed-method study using simulation to detect patient deterioration, Bogossian et al., (2014), concluded that despite pre-briefing the purpose of the simulation, final year nursing students (n=97) lacked the knowledge, skills and teamwork to competently manage a patient whose condition was deteriorating. The researchers recommended curricular changes to include regular rehearsal of situations involving deteriorating patients and expansion of efforts to incorporate simulation throughout the curriculum.

Skills Acquisition

Deliberate practice and simulation

Deliberate practice is a concept described by Ericsson (Ericsson, Krampe, & Tesch-Romer, 1993) whereby an individual employs methods to improve attainment of skill and ultimately to achieve expert level performance. Ericcson developed a theoretical framework to explain the ways in which different levels of performance are achieved as part of deliberate practice (Ericsson, Krampe, & Tesch-Romer, 1993). In his description of the framework, Ericsson details the elements of deliberate practice: 1) the requirement of available time and energy, access to teachers and training resources (resource constraint); 2) the belief in a motivational constraint, or the lack of inherent motivation; and 3) deliberate practice is an activity that can be sustained only for limited time periods (effort constraint). A summary of the concept is given by this definition:

> Deliberate practice involves engaging highly motivated learners in well-defined tasks that are representative of the real world of clinical practice, at an applicable level of difficulty, where informative feedback is promptly available." (Ericsson, Kramper & Tesche-Romer, p. 367).

The framework of deliberate practice is being adopted as a possible framework for use with simulation and in teaching psychomotor skills in nursing (Chee, 2013; Oermann, et al., 2011; Whyte & Cormier, 2014). In the Whyte and Cormier (2013) study, a randomized control design was used with the 20 students in the intervention group (n=-40). The intervention consisted of a deliberate practice educational intervention to determine nurses' ability to provide favorable clinical outcomes for clinically unstable patients in a critical care unit. The intervention consisted of participation in a clinical simulation scenario to care for a patient in a critical care unit, while providing a concurrent verbal report. The student initially participated in the scenario to care for a patient; then repeated the scenario three times, including verbal reports, reflection, an opportunity to access information about clinical practice guidelines, and a final opportunity to run through the complete scenario again. The deliberate-practice intervention group achieved statistically significant improvements in their level of proficiency in caring for the patient.

Oermann, et al. (2011) conducted a randomized control trial (n=606) to explore the effects of using deliberate practice on retention of cardiopulmonary resuscitation (CPR) psychomotor skills. Nursing students engaged in brief, regular practice sessions over the period of one year. Students in the intervention group, who practiced skills regularly on manikins, which provided automated feedback to guide performance, either maintained or improved their level of performance, and performed better than students in the control group.

Pattern recognition

Pattern recognition is described by Benner (1984) and Dreyfus & Dreyfus (1986) as aspects or attributes of a discipline that are learned and used by the novice. These aspects require prior experience in actual situations in which the novice begins to recognize familiar elements that lead to the acquisition of skills and knowledge. Ericsson, Whyte, and Ward (2007) describes how future experts need to seek out particular kinds of experience, which are

deliberate in nature and designed to allow opportunities to reach and exceed performance goals through repetition and problem-solving. Teachers play an integral role by providing immediate feedback which facilitates learning and achievement of increased levels of skill and performance. Christensen (2010) discussed pattern recognition in terms of the ways nurses advance their practice. He states that knowledge which is required to advance practice integrates theoretical as well as practical knowledge, and includes pattern recognition as a way to interpret and make sense of complex clinical situations.

Simulation and Evaluation in Nursing Education

Simulation in nursing education

Simulation is used in nursing education as a teaching strategy and a method for evaluation of student performance. Simulation-based education has been shown to be an effective strategy to teach nursing students to recognize and respond to patients' deteriorating condition (Cooper et al., 2015). Simulation allows students to interact with simulated patients in a safe environment that mimics an actual clinical setting, while allowing students the opportunity to practice without risk to the patient. Dieckmann, Gaba, and Rall (2007) described the ways in which simulated learning environments can be adapted to meet the learning needs of the student and to provide a high degree of realism. Simulation was described to include face to face experiences involving students with high-fidelity simulation manikins, standardized patients and/or virtual simulations through computer-based technology (Cant & Cooper, 2014; Cooper et al., 2010; Cooper et al., 2015, Lamont & Brunero, 2013; Lamont & Brunero, 2014).

Simulation is an appropriate strategy to use for education and evaluation in light of higher patient acuity levels in the hospital, and shorter hospital stays. Simulation provides another layer of evaluation to complement written examinations, which may not sufficiently represent nurses' knowledge and skills in a clinical setting (Wolf et al., 2011). The most important element of decision-making is correct identification and response to a patient problem. The ability to evaluate novice nurses and student nurses' abilities to problem-solve can be enhanced by simulation. Fero's research with nursing students (Fero, et al., 2010) (n=36), although concluding students

experienced difficulty meeting expectations in simulated clinical scenarios, did indicate that high-fidelity human simulation performance was positively correlated with critical thinking scores. Gates, Parr and Hughen (2012) used high-fidelity simulation with undergraduate nursing students (n=104), and found that students who participated in simulation scored significantly higher on course examinations than students who did not participate in simulation. Merriman, Stayt and Ricketts (2014) used simulation and OSCE performance to evaluate the effectiveness of clinical simulation compared to classroom teaching in assessment of the deteriorating patient (n=34). The researchers concluded that clinical simulation was a more effective teaching strategy than classroom teaching in regards to evaluation of a deteriorating patient. A simulation-based educational program was developed for use with nursing students to assist them to achieve competency in assessing patients experiencing physical deterioration (Liaw, Rethans, Scherpbier, & Piyanee, 2011). Researchers examined the program, Rescuing a Patient in Deteriorating Situations (RAPIDS) using a randomized controlled study (n=31 third year nursing students). Clinical performance post-test mean scores improved significantly following the educational program. (Liaw et al., 2011).

Researchers in Australia have conducted numerous studies to evaluate their educational model, called FIRST2ACT (Feedback Incorporating Review and Simulation Techniques to Act on Clinical Trends) (Bogossian et al., 2014; Buykx et al., 2011; Buykx et al., 2012; Cooper et al., 2013; Cooper et al., 2015, Endacott et al., 2010; Endacott et al., 2011; McKenna et al., 2014). According to the model, in addition to initial development of core knowledge and examinations, learners participated in high-fidelity simulation, viewed a video-recording of their performance for the purpose of reflective self-review, and received performance feedback from the clinical expert (Buykx et al., 2012). Researchers conducted studies in three Australian universities with participants including undergraduate nursing students, postgraduate midwifery students, and registered nurses in rural hospital settings. In the early studies of final year undergraduate nursing students (Cooper et al., 2010; Cooper et al., 2011; Endacott et al., 2010), nursing students demonstrated deficiencies in their abilities to respond to patient deterioration; however

the researchers continued to refine the model, substituting standardized patients for manikin-based scenarios. In a study of registered nurses in a rural hospital (Cooper et al., 2011) nurses were unable to recognize deteriorating patient conditions in a timely way during simulation. After the intervention, the researchers reviewed 258 patient records and found nurses were more likely to record observations of patient condition, report pain scores, and intervene to apply oxygen therapy correctly.

Evaluation of clinical performance

Faculty nurse educators participate in evaluation of student nurses, both in didactic performance, and in performance in the clinical setting (Oermann, Yarbrough, et al., 2009a; Tanicala, Scheffer & Roberts, 2011). In evaluating clinical performance, educators must define the clinical behaviors that constitute safe practice and use valid and reliable clinical evaluation tools to minimize or eliminate subjectivity and inconsistency (Gallant, MacDonald, & Smith Higuchi, 2006; Tanicala, Scheffer & Roberts, 2011). Isaacson & Stacy (2009) discuss the need to use rubrics to provide more objective measures for clinical evaluation. In the Tanicala, et al (2011) phase 1 study, focus groups were convened to identify faculty perspectives regarding nursing student behaviors that result in failure in the clinical setting. Common subthemes in this initial phase included the concern for safety in assuring clinical behaviors, critical thinking, communication, and ethical behaviors.

In the Evaluation of Learning Advisory Council of the National League for Nursing Study (Oermann, Saewert, Charasika, & Yarbrough, 2009a; Oermann, Yarbrough, Saewert, & Charasika, 2009b), nursing faculty were surveyed about their practices of grading students' clinical performance. Faculty outlined expectations used to evaluate student clinical performance which included ability to analyze complex patient situations, ability to critically think and solve problems, ability to communicate effectively with all members of the health team and clients receiving care, and characteristics of effective leadership, and cultural and spiritual competency. Faculty responding to the survey (n=1,573) identified challenges in evaluating clinical performance which were based on the tool, faculty role, and role in assigning a failing grade.

Faculty identified the need for valid and reliable tools that would yield consistent ratings based on multiple raters. Faculty characteristics that caused concern were when faculty served as both teacher and evaluator. Finally, faculty discussed their reluctance in evaluating students' poor performance that results in clinical failure.

Clinical failure and failure to fail

Although educators are endeavoring to define and delineate clinical failing behaviors, they must address the concern about reluctance or refusal to allow student nurses to fail, when their behaviors do not meet assessment standards. The concept of "failure to fail" is another phenomenon that supports the need for valid and reliable evaluation tools. "Failure to fail" students arises when faculty assign a passing grade when in fact, the student has failed to achieve a passing standard (Black, Curzio & Terry, 2014; Danyluk, Luhanga, Gwekwerere, MacEwan & Larocque, 2015; Docherty & Dieckmann, 2015; Vinales, 2015). In a nursing study, Docherty & Dieckmann (2015) surveyed faculty from community colleges and universities in a state using a shared nursing curriculum, to explore the extent to which this phenomena occurred, and the possible reasons. Faculty respondents (n=84) responded to a 37-item survey tool which included items related to their past experience with failing students; whether they had previously failed a student clinically; whether they had given students the benefit of the doubt when determining clinical competency, and whether they had ever awarded a higher grade than they felt was deserved. Seventy two percent (72%) indicated they had given the student the benefit of the doubt and 43% reported they had awarded higher grades than they believed had been earned. The survey also included reasons for failing to fail, such as faculty knowing the student, being concerned about administrative support, and fearing litigation. As part of the discussion of findings, Docherty & Dieckmann (2015) included the practice of using team-grading norms and the need to use grading rubrics that had been subjected to rigorous testing. Without valid and reliable tools, faculty believed that grading rubrics could potentiate grade inflation or result in a greater incidence of failing to fail a student.

Two qualitative studies addressed the issue of failure to fail students in clinical practice (Danyluk, et al., 2015; and Black, Curzio & Terry, 2014). Danyluk, et al., (2015) explored the failure to fail in evaluating teacher candidates in Canada. The researchers interviewed university supervisors and associate teachers who had experience with students at risk of failing. The study yielded several themes for reasons educators (n=6 educators) failed to allow a student to fail their clinical practicum including: 1) the difficulty involved in the process of failing a student; 2) the potential impact of failure, including failing to graduate; 3) the additional work for faculty related to failing a student; and 4) the potential negative consequences for the program or university of student failure. In the Black, Curzio and Terry study (2014), nurse midwife mentors involved in supervising and evaluating students' clinical performance participated in a hermeneutic study. Themes from this study included the experience of moral stress, demonstration of moral integrity and the need to ensure moral residue which they defined as the values that remained despite needing to make difficult decisions.

Grade inflation

Grade inflation is defined as a greater percentage of high or excellent scores than is warranted by students' performance (Paskausky & Simonelli, 2014; Donaldson & Gray, 2011; Isaacson & Stacy, 2009; Moskal & Leydens, 2000). Some of the reasons attributed to grade evaluation can be categorized in terms of students, faculty evaluators, the relationship between student and faculty, and the grading tool (Donaldson & Gray, 2011). Moskal and Leydens (2000) note that well-constructed tools can improve intra- and inter-rater reliability. Isaacson & Stacy (2009) note that face-to-face evaluations of clinical performance may lead to evaluator leniency, in order for faculty to avoid conflict. Paskausky & Simonelli (2014) examined the relationship between licensure-style written final course examinations and faculty assigned clinical grades in undergraduate nursing students in a maternal nursing course. Their study findings indicated a 98% clinical grade discrepancy score indicating likely grade inflation. More concerning than the notion of inconsistency in clinical grading was the potential for underperforming or unsafe students progressing through the program, passing licensure exams, and potentially be providing

unsafe care to patients. This study did not address reasons for grade inflation, but speculated that faculty aversion to discomfort with assigning failing grades could be a concern.

Simulation evaluation tools

The literature abounds with papers and studies regarding tools used to evaluate student performance during simulations not involving an OSCE. Typically OSCEs involve evaluation of individual student performances during a high-stakes assessment; whereas simulation frequently involves students working in groups. Kardong-Edgren, Adamson, and Fitzgerald (2010) published a review of simulation evaluation instruments in which they categorized the instruments by purpose, learning domain employed, and whether analysis of reliability and validity were reported. The tools used varied from paper and pencil test to scoring tools with rubrics. Of the eleven instruments, only the Simulation Evaluation Instrument (Todd, Manz, Hawkins, Parsons, & Hercinger, 2008) reported interrater reliability which was .85 to .89.

A subsequent study (Adamson & Kardon-Edgren, 2012) compared three simulation instruments for rater reliability (n=38 raters): the Lasater Clinical Judgment Rubric (Lasater, 2007); the Seattle University Evaluation Tool (Cicero & Mikasa, 2008); and the Creighton Competency Evaluation Tool (Todd, et al., 2008). These 3 tools were evaluated by thirty-eight nurse educators who viewed National League of Nursing (NLN) developed scenarios performed by students at 3 levels of performance. Rater reliability, internal consistency and test-retest reliability were evaluated. The interrater reliability using intraclass coefficients was .889 for the Lasater Clinical Judgment Rubric; .858 for the Seattle University Evaluation Tool; and .952 for the Creighton Simulation Evaluation Instrument. The Creighton Simulation Evaluation Instrument was modified for use in the National Council of State Boards of Nursing (NCSBN) Simulation study and renamed the Creighton Competency Evaluation Tool (CCEI) (Hayden, Keegan, Kardong-Edgren, & Smiley, 2014a). The CCEI was used in an NLN-funded project to evaluate the process of using simulation for high-stakes evaluation (Rizzolo, Kardong-Edgren, Oermann, & Jeffries, 2015). The project involved scenario development, recruitment and training of expert raters, pilot testing followed by field testing in schools of nursing (n=28 videos). Nine schools of nursing

participated by having students complete two scenarios. Scenarios were video-recorded and viewed by eleven expert raters. Kappa and intraclass correlation coefficient (ICC) scores were computed and were determined to fall within the “good” range with interrater agreement, kappa 0.65 for time 1 and 0.66 for time 2; ICC was 0.65. When testing the subscales of the CCEI, intrarater reliability remained “good” but interrater reliability was determined to be “fair” with a range of 0.42 to 0.48.

OSCE and OSCE evaluation tools

The earliest literature describing an OSCE was in the medical literature, defining an OSCE as a measure used to evaluate competency of British medical students nearing completion of their program (Harden, Stevenson, Wilson Downie, and Wilson, 1975). The OSCE evaluation consisted of sixteen (16) stations, some of which employed real patients or patient actors. Faculty evaluators were assigned one to a station; therefore, students were evaluated by up to 16 different faculty members (Harden, et al., 1975).

A review of the nursing OSCE literature included theoretical and practical articles describing the process and conduct of an OSCE, development of scenarios, and appropriate evaluation statistics. Benefits of OSCEs reported in review literature include that they represented a more objective measure than many other areas of practice (Schuwirth & van der Vleuten, 2003; Watson, Stimpson, Topping, and Porock, 2002); allowed a broad range of skills to be tested (Watson, et al., 2002); had a high level of reliability and validity (Schuwirth & van der Vleuten, 2003; Bartfay, Romborough, Howse, & Leblanc, 2004); and included strong motivation for learning (Bartfay, et al., 2004). In undergraduate education, OSCEs can be used most effectively in conjunction with measurable aspects of performance in a clinical setting including technical skills, critical thinking, interpersonal and communication skills (Mitchell, Henderson, Groves, Dalton, & Nulty, 2009). In the 2007 NLN Survey conducted by the Evaluation of Learning Advisory Council (Oermann, et al., 2009a) nursing faculty were surveyed to determine the types of assessment and evaluation strategies they used to evaluate the cognitive and affective

domains. A total of seventy-seven respondents reported using an OSCE to evaluate the affective domain.

McWilliam and Botwinski (2010) describe the development of a nursing OSCE including development of scenarios, training of standardized patients and student perceptions of the OSCE experience. The purpose of their study was to examine key aspects of a nursing OSCE used to develop a performance-based evaluation of student competencies (n=60 nursing students, three faculty members and three standardized patients). Findings from the study were descriptions of the process for development of the scenarios, role and training of standardized patients, and student perceptions of the experience.

Few research studies were found about OSCE rating tools to evaluate performance of students or practicing nurses. Studies reviewed included development of an OSCE tool using factor analysis (Walsh, Bailey, Mossey & Koren, 2010) (n=565 first-term students across 4 nursing programs). This study provides an OSCE checklist with 5 elements, evaluated by 5 behaviors using a Likert scale, with no rubric to describe the specifics of each rating of the scale. No results were presented on the findings of student performance. Rentschler, et al. (2007) conducted a pilot study to evaluate an OSCE for senior undergraduate nursing students. The purpose of this study was to evaluate effectiveness of an OSCE designed to evaluate clinical competencies of senior baccalaureate nursing students (n=49). Students interacted with a standardized patient to perform an assessment and interventions. Evaluation was performed by the standardized patient completing a checklist of skills performed, and the student's post-encounter form to evaluate satisfaction with the experience.

The FIRST2ACT study used a standardized OSCE score sheet. The score sheet included 24-28 items, based on the scenario used. Oranye, Ahmad, C., Ahmad, N., and Bakar (2012) conducted research to evaluate performance of practicing nurses enrolled in open distance learning courses using a five-station OSCE (n=569 students). The researchers evaluated clinical skills competence in 14 learning centers, and factor analysis to design the study tool. The methodology is described as rating students' clinical competency in five skills stations,

in six areas of specialization using standardized questions scored by experienced faculty raters. The checklists were not included in the published study. Only 14% of participants demonstrated the highest level of competency, despite being experienced, practicing nurses.

In a recent study describing use of an OSCE, Raurell-Torreda et al. (2015) used scores on an OSCE with a human patient simulator and National League for Nursing (NLN) validated case scenarios to compare skills acquisition in the two groups of undergraduate nursing students (n=101, with 35 assigned to the intervention group). Students in the intervention group developed better patient assessment skills than students in the control group. The OSCE checklist they used was developed and published as a guideline to evaluate student progress toward the objectives of a clinical course (Wolf et al., 2011). The authors included their OSCE instrument in the published study. Interestingly, both Raurell-Torreda, et al. (2015) and Wolf et al. (2011) mentioned that early versions of that same OSCE Checklist had been validated with a reported interrater reliability of 95% in Henneman, Cunningham, Roche, & Curnin, (2007), however findings from that study did not include rater reliability of an instrument.

Two studies provided evidence that the use of an OSCE corresponds to improved performance (Merriman, Stayt, & Ricketts, 2014; Preston-Safarz & Bolick, 2015). In the Merriman et al. study (2014) a pilot project was conducted to determine whether clinical simulation was more effective than traditional classroom methods in teaching assessment skills necessary to recognize and respond to the deteriorating patient (n=34 students, with 15 students in the intervention group). Both methods included instruction on the ABCDE assessment method, but the intervention group included a 2-hour simulation. The OSCE checklist consisted of 24 objective performance criteria. Students in the intervention group performed better on post-intervention OSCEs than the group that did not receive simulation. In the RN to BSN study (Preston-Safarz & Bolick, 2015), the purpose was to determine whether using simulation and an OSCE would provide faculty members the opportunity to evaluate skill attainment and transfer of student learning (n=6). Results included demonstration of mastery of course content and high faculty satisfaction.

A recent U.S. study was conducted in 2014 (Stiller, et al., 2015) to refine an OSCE rating tool for comparison with a previously existing tool used in an educational setting. Three versions of the original tool existed corresponding with three clinical vignettes (pyelonephritis, heart failure and asthma), and each contained weighted items. Faculty members used the tool by assigning values for each weighted item during evaluation of student performance. The investigators revised the items in the tools to be scored as done or not done in order to provide a more objective measure of each element of the tools. The total score for the revised tools was based on a 100-percentage-point system. The revised tools were evaluated by use by course faculty members (n=76 new tool; n=49 original tool). Reliability testing of original and revised versions of the tool showed improved reliability of the new instruments (Cronbach's alpha of the original pyelonephritis tool: 0.22; revised tool 0.68; original asthma tool 0.24; revised tool 0.62.

Nursing faculty members need reliable and valid tools to evaluate student performance in simulation scenarios using OSCEs in high-stakes testing (Bensfield, Olech, & Horsley, 2012). These investigators used high-stakes evaluation and simulation with senior nursing students to observe students' performance. They used the findings from their data collection to make curricular revisions.

Reliability

Rubrics are essential tools for use in evaluation of clinical performance. Criterion-referenced rubrics include detailed performance criteria which delineates student expectations, across levels of performance (Walshe, O'Brien, Hartigan, Murphy, & Graham, 2014). To avoid subjectivity, rubrics must be evaluated for intra-rater and inter-rater reliability. Inter-rater reliability must be established for each new population (Adamson & Prion, 2012).

Reliability is an important characteristic of an assessment tool and is essential for assuring rigorous study outcomes (Lim, Palethorpe, & Rodger, 2012). Intra-rater reliability refers to the degree of consistency of an individual rater over time (Walshe, et al., 2014). Inter-rater reliability is described as the level of agreement between multiple raters using the same instrument on the same set of subjects (Lim, Palethorpe, & Rodger; Walshe, O'Brien, et al.,

2014). Establishment of inter-rater reliability suggests that the tool used to measure a particular phenomenon is stable and consistent when used by multiple raters. A rater is defined as a person who conducts biophysical measurements with specific biomedical devices or equipment, including tests or as one who observes or judges in a study (Ryan-Wenger, 2010; Grove, Gray & Burns, 2015).

Lim, Palethorpe, and Rodger (2012) describe different types of inter-rater reliability and the statistics used in measurement. In studies in which scoring rubrics or rating scales are used consensus estimates are the appropriate analysis to establish the degree of agreement between raters. Percentage agreement and Cohen's Kappa are statistics used in consensus estimates.

Grading rubrics and instruments used to measure student performance in simulation are being developed and tested to establish reliability. Methods employed in reliability testing include recruitment of educators possessing defined characteristics and experience with simulation; multi-site testing of instruments; use of video-recordings allowing educators in multiple sites to view and rate performances; and use of test-retest measures where raters viewed video-recorded vignettes and rated performance, then at another point in time, viewed and rated the same performance a second time (Adamson & Kardong-Edgren, 2012; Adamson, Parsons, et al., 2011; Cordi, et al., 2012; Parsons, et al., 2012; Walshe, et al., 2014).

The need for reliable and valid simulation and OSCE evaluation instruments. The literature on OSCEs, while extolling the worth and value of this type evaluation, is specific in documenting the need for evidence of reliability and validity for evaluating student performance (Cazzell & Howe, 2012; Mitchell et al., 2009; Rushforth, 2007). Few instruments have undergone testing to establish reliability and validity prior to implementing their use in practice (Hayden, Keegan, Kardong-Edgren, & Smiley, 2014a; Kardong-Edgren, Adamson, & Fitzgerald, 2010; Adamson, Kardong-Edgren, & Willhaus, 2013; Cordi, et al., 2012). In their updated review of simulation instruments, Adamson, Kardon-Edgren & Willhaus (2013) includes data from the Egyptian study (Selim, Ramadan, el-Gueneidy, & Gaaafer, 2012) which detailed an 11-station nursing OSCE used to evaluate psychiatric nursing students' performance. This study is the only

simulation study they reviewed which includes findings from an OSCE, but whether this multi-faceted OSCE could be used for acute care nursing performance is uncertain.

Kardong-Edgren et al. (2010) makes the case that researchers, in order to add to the body of research in the effectiveness of human patient simulation, must develop reliable and valid tools to evaluate student performance in simulation. Rather than developing new tools, educators should conduct research to establish reliability and validity on existing tools. Adamson et al. (2013) stress that tool selection must go beyond reliability and validity to consider whether the tool appropriately measures the activity and behaviors for which it is purported to be used.

In the Standards of Best Practice developed by the International Nursing Association for Clinical Simulation (INACSL) (2011), the group identified criteria for educators to use when evaluating student performance in order to achieve valid and reliable results, including the need to use standardized scoring methods. In Standard VII: Participant Assessment and Evaluation (Sando, Meakim, Gloe, Decker & Borum, 2013), the guideline statement emphasizes that interrater objectivity and reliability are increased by the use of standardized checklists which focus on assessment of specific skills.

OSCE evaluation tools have psychometric characteristics that indicate a tool has the capacity to be able to measure the constructs intended to be measured and do so in a consistent manner. Further, Cordi, et al. (2012) states that reliability tests must be conducted on new tools because of the knowledge that there is an inherent degree of random error in any measurement instrument. In their review of objective measures, Cant, McKenna and Cooper (2013) reported on a number of studies that assessed pre-registration nursing students' clinical competence. The research reviewed included a study conducted to evaluate self-efficacy and perceived competence (Lauder, et al., 2008) and a study involving psychiatric students in Egypt using 11 OSCE stations, with no reported reliability or validity data (Eldair, et al, 2010). Nursing faculty need to either use tools which already have established OSCE performance checklists and/or with demonstrated reliability and validity.

Retrospective Secondary Data Analyses

Retrospective studies have been used in nursing education to examine cohorts of students in simulation and during clinical performance. A retrospective observational study was conducted to investigate the role in predicting nursing student academic success by using upper-secondary diploma grades and a score on a program admission test (Lancia, Petrucci, Giorgi, Dante & Cifone, 2013). Data was obtained including five cohorts of nursing students in an Italian bachelor's degree in nursing program. The researchers concluded that there was a correlation between failure from the program and low grades from the upper-secondary diploma course work, and it was statistically significant. Further the researchers concluded that upper-secondary diploma coursework grades should be given consideration when planning for increasing enrollment and admission of students.

Hall (2015) conducted a retrospective study to examine the effectiveness of using high-fidelity simulation with senior students enrolled in a maternity course in a baccalaureate program. The investigators explored whether students who received instruction through the use of high-fidelity simulation added to traditional hospital-based clinical instruction achieved greater critical thinking skills and higher NCLEX scores than students who received traditional clinical instruction alone, as measured by a standardized, content-specific exam. Simulation students scored significantly better on NCLEX scores than did non-simulation students. Investigators surmised that using high-fidelity simulation as an instructional measure enhanced confidence in clinical performance, which then translated into higher NCLEX score

In one additional retrospective descriptive study, a review of records was used to explore medication errors made by undergraduate nursing students, and factors believed to contribute to the errors. The secondary data analysis indicated that seven students out of 325 had reported making a medication error (Tabassum, Saeed, Dias, & Allana, 2016). Medication errors included unsupervised preparation of the medication, failure to correctly identify the patient, using the wrong route and the wrong dose of medication. Factors contributing to the medication errors included increased patient acuity, insistence by staff and stress due to environmental factors on

the unit. Investigators planned to incorporate strategies for medication error prevention in future cohorts of nursing students.

Summary

Simulation has been used extensively in health education, with incorporation into nursing education becoming prevalent over the last 10-12 years. Simulation allows nurses to practice and interact in clinical situations in a safe environment, where learning can occur. Coupling simulation with OSCEs is an effective strategy to objectively measure student performance in the psychomotor and affective domains. Nurse practice acts and clinical standards of professional nursing organizations recognize the essential nature of nurses' skill to perform a comprehensive and/or focused assessment of the patient, as a baseline for then being able to recognize when a change or deterioration in condition occurs. Novice nurses have had fewer opportunities for mentoring in the new professional environment, and often lack skills necessary to allow them to respond to the changing patient condition in a timely manner, to assure positive patient outcomes. Nurse educators have developed OSCE scenarios, including skills they deem essential to measure a set of core competencies. OSCE rating tools can be used to delineate expectations and for ease of use by raters.

There is a lack of OSCE rating instruments that have been tested to establish that they are reliable and valid to be used in evaluating students in summative evaluations. Many simulation tools have been developed to evaluate various aspects of simulation, but none have been identified that meet the goals of evaluating nursing performance in a high-stakes simulation OSCE. Without rigorous testing, educators have no way of establishing that their evaluation of student performance is effective and equitable. This study will be conducted using a faculty developed OSCE rating tool to describe its use and to establish inter-rater reliability, in order to add to the body of nursing knowledge.

Chapter 3

Methods and Procedures

Introduction

This retrospective research study was conducted for the purpose of determining inter-rater reliability of the OSCE rating tool when used by internal and external faculty raters. The original study was performed to examine psychometric properties of a faculty-developed OSCE rating tool. (Stiller, et al., 2015). This chapter includes the research design, setting and procedures of this retrospective, secondary data analysis.

Research Design

The research design of this study was a retrospective, descriptive study using secondary data to determine whether the OSCE rating tool was reliable when used by multiple course faculty raters and faculty raters not assigned to the course. The tool was developed to evaluate performance of senior nursing students in a capstone course during an OSCE using standardized, simulated clinical scenarios.

Retrospective studies are those in which researchers use previously collected data to answer a current research question. They may be classified as cohort or case-control designs (Abbott, Barton, Terhorst & Shembel, 2016). Cohort studies examine a group of subjects with either a similar condition or those who have experienced a specific event (Grove, Gray & Burns, 2015). Retrospective studies can either be designed as experimental or observational (Abbott, et al., 2016). In the current study, data from a prior study were analyzed, where two groups of faculty rated student performances during an OSCE. Concerns about using secondary data include methodological problems of the study during which the data were collected, such as the possibility of less than rigorous data collection, and knowledge problems, such lack of clarity about the sample and data analysis procedures (Law, 2005). Advantages of using secondary data are that it provides an inexpensive method to quickly answer research questions (Abbott, et al., 2016).

Nurse researchers have effectively used secondary data analysis in prior studies on the topics of factors predicting success of students in nursing education (Lancia et al., 2013) and of factors influencing medication errors by nursing students (Tabassum et al., 2016). Specifically related to simulation, Hall (2015) used secondary data analysis to compare NCLEX pass rates of students receiving traditional clinical instruction plus simulation with the pass rates of students receiving traditional clinical instruction.

This design was selected because of the availability of data collected using an OSCE instrument to evaluate student performance using the variables of Patient Safety, Assessment, Planning and Medication. When designing a study using previously collected data, researchers must ensure that the data are appropriate to answer the research questions (Doolan & Froelicher, 2009; Waltz, Strickland & Lenz, 2005). The data collected using the revised OSCE instrument were consistent with the purpose of this current study and appropriate to answer the research questions. Data from the prior study were not analyzed to examine differences between Course Faculty raters and Non-course Faculty raters. This gap formed the basis of the research questions for the current study.

Description of Prior Study

Sample

The sample of the prior study (Stiller, et al., 2015) consisted of six Course Faculty from the college of nursing and three (3) faculty raters not assigned to the course. Two of the faculty raters not assigned to the course were from the college of nursing and were selected because they had experience with simulation or simulator research. One faculty rater was a Doctor of Philosophy in Nursing student who was a graduate research assistant in the college of nursing, and was interested in simulation. Forty-four (44) student performances during a summative OSCE using simulation were evaluated by the Course and Non-course faculty raters.

Setting

The setting for this study was a college of nursing in a state university with a state-of-the-art clinical simulation laboratory in a large baccalaureate of science in nursing (BSN) program in

the Southwestern United States. Faculty members had numerous opportunities to engage in simulation activities including use of high-fidelity simulation manikins and human simulators (standardized patients) to conduct simulated scenarios and OSCEs.

Background of the OSCE

Students in the senior semester of the nursing program were enrolled in the final clinical course. As part of the preparation for the individual student's clinical experience, students participated in an OSCE evaluation using clinical simulation to determine their ability to perform a focused assessment and recognize significant clinical findings or deterioration in the patient's condition.

Faculty members worked with the clinical simulation specialists, who prepared the simulation laboratory using one of three clinical vignettes involving simulated patients with 1) congestive heart failure exacerbated by pneumonia; 2) respiratory distress due to asthma; and 3) pyelonephritis with decreased urine output and elevated potassium level. These vignettes were developed by the National League for Nursing (NLN) and validated for accuracy. The clinical course faculty adapted these scenarios and used them for the clinical vignettes in the OSCE evaluation.

The OSCE procedure included the student entering the simulated hospital room and being briefed by the Course faculty member using standardized instructions. The Course faculty provided the student with a patient chart, including physician orders, lab work, Medication Administration Record, and a current nursing shift report with patient chief complaint. The faculty member instructed the student that the simulated patient would experience a symptom during the scenario. The patient's vital signs were displayed on a monitor. A simulation specialist spoke for the patient, and the simulation manikin was programmed to have chest movements with breathing, possibly a cough, or other voice sounds consistent with the clinical vignette.

The student was instructed to interact with the patient (voiced by the simulation specialist) to perform a focused assessment, recognize a patient problem, appropriately intervene, correctly administer medications, and maintain patient safety. Each student was given

30 minutes to complete the patient vignette and 30 minutes for debriefing following the performance. The faculty member's role was to remain silent during the student's performance, rate the student's performance using the revised OSCE Instrument, and then debrief the student at the bedside to review the student's performance. Students' performances were video-recorded during the OSCE and those recordings were used by the Non-course Faculty raters to assign ratings.

Development and Revision of OSCE Instrument Used for Data Collection

In the prior study, (Stiller, et al., 2015) faculty took the three instruments previously developed for use in evaluation of student performance in the clinical course. The original instruments were three scenario-specific instruments corresponding to the clinical vignettes and consisted of items of varying weights. Faculty revised these weighted instruments to be checklist instruments consisting of twenty-eight (28) items reflecting actions to be performed under the categories of Patient Safety, Assessment, Planning and Medication. Each item had a point value of one (1) for "Done" and zero (0) for "Not Done". See the revised OSCE instrument in Appendix A.

Patient safety was evaluated by the nurse identifying self to patient; identification of the patient using two identifiers; and performing hand hygiene on entry and exit from the room. Assessment and nursing diagnosis were evaluated in each scenario by observing performance of vital signs, assessment of intravenous (IV) site, and a focused assessment specific to the scenario. The student was expected to recognize a variation from normal in patient condition and/or laboratory or diagnostic findings. Interventions and re-assessment were evaluated by the student performing interventions to improve the patient condition, followed by reassessment. Medication administration was evaluated by observance of the five rights, correct identification of patient allergies, and administration of scenario-specific medications administered by intravenous (IV) Push, using correct technique. The OSCE evaluation required students to achieve 85% on the items in Patient Safety, Assessment and Planning and 100% on Medication.

Institutional Review Board Approval

The researchers submitted an application to the Institutional Review Board (IRB) at the university where the study was conducted. The IRB determined that approval was not needed because the study purpose was instrument development.

Informed Consent

Senior nursing students within six weeks of graduation were given information about the study. These students had previously given consent to have their simulation experiences video recorded, and were asked to consent to allow their video recorded OSCE performances to be included in this study. A consent form was provided and students indicated their consent by signing the form.

Ethical Considerations

OSCE performances were video-recorded and labeled by code number and no student names were included. The recording equipment was housed in a separate room and was not obvious to the students. Recordings were kept in a secure location. Decisions regarding pass/fail of the OSCE performance were determined strictly by the Course faculty using the original OSCE instrument.

Risks to students involved in the study could potentially have been the impact of having their OSCE performance video-recorded. The recording of performances could cause an increase in anxiety and affect the student's performance. Potential benefits to students would be the benefit to future students from availability of a reliable tool for evaluating OSCE performance.

Data Collection and Results

The six Course faculty members used a combination of original and revised OSCE instruments, which corresponded to the clinical vignette, to rate the student's OSCE performance. The three Non-course faculty viewed the recordings of the same student OSCE performances using the same instrument as the Course faculty. The Non-Course faculty raters watched a live stream performance or viewed recorded videos and scored the performances. No rater training occurred prior to the study to evaluate the revised OSCE instrument.

Ratings were collected on a cohort of approximately 80 students during a recent senior semester. The ratings of the six Course faculty and the three Non-Course faculty reviewers were entered into a statistical software program, IBM SPSS Statistics, version 23 (2015) for analysis. Data were coded by Student number, Course faculty evaluator and Non-Course faculty evaluator, and by clinical scenario. Ratings were entered as the score earned for each of the four categories. Forty-four (44) students were evaluated using the revised OSCE instrument.

The faculty reported Cronbach's alpha for both the original and revised OSCE instruments. For the OSCE instrument specific to asthma, the original Cronbach's alpha was reported as .24; the revised Instrument was reported as .62. For the pyelonephritis scenario, the original instrument was reported to have a Cronbach's alpha of .22 and the revised instrument Cronbach's alpha was reported as .68. The sample was too small to calculate Cronbach's alpha for the heart failure instrument.

Current Study Using Secondary Data Analysis

Ethical Considerations

The researcher completed application requirements to obtain IRB approval from the university where the study was conducted. The IRB determined that the study met the requirements for exempt status since the data was obtained from an existing database, and no identifying subject information was included.

Measurement Methods and Data Collection

Data were collected as part of the prior study and entered into the database. Analysis of data is described below.

Data Analysis

Analysis of data by statistical method is summarized in Table 3-1 below. The specific statistic and related analysis are included.

Table 3-1 Statistical Methods

Statistics	Analysis
Descriptive statistics	Range of scores, Mean, Median of scores for Patient Safety, Assessment, Planning and Summary Score
Kruskal-Wallis and Chi-Square	Differences between Course Faculty's ratings; Differences between Non-course Faculty's ratings for Patient Safety, Assessment, Planning and Summary Score
Wilcoxon Signed Ranks Test	Differences between Rater groups for Patient Safety, Assessment, Planning, and Summary Score
McNemar Test	Differences between Rater groups on Medication
Spearman Rho	Association/Agreement between Rater groups on Patient Safety, Assessment, Planning and Summary Score
Kappa statistic	Agreement between Rater groups on OSCE pass/fail

Descriptive statistics

IBM SPSS Statistics 23 was used to perform analysis of the data from this database. Description of the demographics of the study sample included frequencies of student performances rated by each of the six course faculty; frequencies of student performances rated by each of the three Non-Course faculty reviewers, and frequencies of scenarios used. Data were assessed for normality. Analysis of data addressing the research questions included descriptive statistics for the range of scores, median, mean, and standard deviation on the variables of Patient Safety, Assessment, Planning and Medication. A new variable, Summary Score, was computed consisting of the total score from Patient Safety, Assessment, and Planning.

Research questions

Research question 1. Were there differences among the six Course Faculty's ratings of student performance using the revised OSCE instrument?

A Kruskal-Wallis test, which is a non-parametric test was used to examine differences between ratings among the Course faculty for the variables of Patient Safety, Assessment,

Planning and Summary Score. A Pearson chi-square test was used to examine differences among Course faculty for Medication.

Research question 2. Were there differences among the three Non-course Faculty's ratings of student performance using the revised OSCE instrument?

A Kruskal-Wallis test was used to examine differences among Non-course faculty raters on the variables of Patient Safety, Assessment, Planning, and Summary Score. A Pearson Chi-Square test was computed for the variable of Medication.

Research question 3. Were there differences between ratings of Course Faculty and Non-course Faculty on student performance using the revised OSCE instrument?

A nonparametric test, Wilcoxon Signed Ranks Test, was used to examine differences between groups of faculty raters (Course faculty and Non-course faculty) on the variables of Patient Safety, Assessment, Planning, and Summary Score. A McNemar nonparametric test was used to examine differences between groups of faculty raters on the categorical variable Medication (pass/fail).

Research question 4. What were the inter-rater reliability coefficients for each subscale on the OSCE instrument when scored by multiple raters?

A Spearman's Rho was computed for the variables of Patient Safety, Assessment, Planning and the computed variable Summary Score to examine agreement and rater reliability on these variables because they do not have a normal distribution, and are continuous variables.

The Summary Score variable was recoded into a binary categorical variable of pass/fail based on the score used by faculty raters to determine successful completion, which was an 85%. Using a Kappa coefficient, analyses were performed to examine whether Course faculty and Non-course faculty raters agreed on passing or failing the performance. Inter-rater reliability is the process of determining consistency of agreement among a group of expert raters to assign scores to the same set of behaviors in the same measurement situation (Waltz, Strickland & Lentz, 2005). Additionally, a Spearman's Rho test was used to evaluate internal consistency on the variables of Patient Safety, Assessment, and Planning, as well as the computed variable of

Summary score. Analyses were conducted to determine whether there was a correlation between these subscales of the revised OSCE instrument.

Delimitations

Data were previously collected and entered into a database as described in this chapter. The research questions for this secondary analysis were designed to be consistent with the available data.

Summary

Nurse educators need reliable and valid tools to evaluate student performance, particularly in the area of recognition of clinical patient deterioration and appropriate response. Newly graduated nurses will be in situations in which they must make clinical decisions about patient care and deteriorating patient conditions. Students as well as practicing registered nurses have been identified as having deficiencies in early recognition and response to patient situations. Students should have practice and evaluation of these skills prior to graduation and entry into professional practice. Clinical simulation can be used in educational institutions to prepare students to manage these complex situations, in a safe environment. These performances are often high-stakes in nature, so it is critical that tools be developed that are evaluated for reliability and validity as they impact the students' ability to progress to graduation and entry into practice, as well as evaluating their ability to safely manage complex clinical situations. Providing a reliable and valid tool will add to the body of nursing knowledge.

Chapter 4

Findings

Introduction

Analysis of data from an existing database was performed to determine whether a newly revised faculty-developed OSCE checklist tool was a reliable tool when used by multiple raters. Descriptive statistics for the ratings of course faculty and external blinded reviewers will be reported. Results of nonparametric tests are included to report differences between faculty rater groups as well as agreement between raters.

Results

Data Description

The dataset included course faculty and non-course faculty ratings of subjects' performances (n=44) during an end-of-course OSCE (Objective Structured Clinical Examination). The data were linked to the four subscales of the revised OSCE instrument, which were patient safety, assessment, planning, and medication administration. A copy of the revised OSCE instrument is found in Appendix A of this document.

Analysis for Normality of Data

Descriptive statistics were computed to determine whether the data were distributed normally. Rating scores for each category and the summary score were not normally distributed, with scores being negatively skewed. For Course faculty, skewness was -2.004 for Patient Safety; -.588 for Assessment; -.348 for Planning and -.430 for Summary Score. Negative skewness values indicate a predominance of high scores (Field, 2009). For Non-course faculty, skewness scores were also negative with Patient Safety -.801; Assessment -.310; Planning -.468, and Summary Score -.582. Because the data do not have a normal distribution, non-parametric statistics were computed for analysis. Descriptive statistics are presented in Table 4-1 below

Table 4-1 Descriptive Statistics for Patient Safety, Assessment, Planning, and Summary Score

	Course Faculty		*Non-Course Faculty*	
	Range	Median Score	Range	Median Score
Patient Safety	18.00-30.00	30	12.00-30.00	30
Assessment	29.40–50.00	45	23.10-50.00	38.9
Planning	5.00-20.00	15	0.00-20.00	10.8
Summary Score	62.60-100.00	88	45.00-91.20	77.5

Research Questions

Research question 1. Were there differences among the six Course Faculty's ratings of student performance using the revised OSCE instrument?

Analysis of differences among course faculty raters was conducted using a Kruskal-Wallis test to compare differences in rating scores on the dependent variables of Patient Safety, Assessment, Planning, and Summary Score between the Course faculty raters. There were no significant differences between Capstone course faculty's ratings of subjects on Assessment (χ^2 = 8.811; p=.117). There were significant differences between Capstone course faculty ratings on Patient Safety (χ^2 = 13.694; p=.018); Planning (χ^2 =17.175; p=.004); and Summary Score (χ^2 =16.038; p=.007). The Medication data could not be analyzed for differences because they failed to meet the assumptions of a Pearson chi-square test due to too few non-zero values. Results are presented below in Table 4-2.

Table 4-2 Analysis of differences among Course faculty ratings

Variables	$\bar{\chi}$	*SD*	*p*
Patient Safety	28.636	2.854	.018
Assessment	43.828	5.551	.117
Planning	14.591	4.956	.004
Summary Score	87.056	10.144	.007

Research question 2. Were there differences among the three Non-course Faculty's ratings of student performance using the revised OSCE instrument?

Analysis of differences between Non-Course faculty raters was performed using a Kruskal-Wallis test to compare rating scores on the dependent variables of Patient Safety,

Assessment, Planning, and Summary Score of Non-Course faculty raters. There were no significant differences between Non-Course Faculty raters' scores on the variables of Patient Safety (χ^2 = 1.330; p=.514); Assessment (χ^2 =.726; p=.695); Planning (χ^2 = 2.275; p=.321); or Summary Score (χ^2 = 1.951; p=.377). For Medication, a Pearson chi-square was computed, showing there were no significant differences between Non-Course faculty raters' ratings of Medication (χ^2 = .733; p= .693). Results are presented below in Table 4-3.

Table 4-3 Analysis of differences among Non-course Faculty Ratings

Variables	$\bar{X}$	*SD*	*p*
Patient Safety	25.727	5.276	.514
Assessment	38.283	6.959	.695
Planning	10.784	4.470	.321
Summary Score	74.794	10.763	.377

Research question 3. Were there differences between ratings of Course faculty and Non-course Faculty on student performance using the revised OSCE instrument?

Differences between two groups of faculty raters (Course Faculty and Non-course Faculty) were analyzed using a Wilcoxon Signed Ranks test. This test was performed to examine whether differences existed between groups of raters using the revised OSCE instrument. There were significant differences between groups of raters on all variables of Patient Safety, Assessment, Planning, and Summary Score. See Table 4-4 below.

Table 4-4 Analysis of Differences Between Rater Groups

	Course Faculty Raters			*Non-Course External Raters*		
	$\bar{X}$	Median	SD	$\bar{X}$	Median	SD
Patient Safety	28.636	30.00	2.854	25.727	30.00	5.276
Assessment	43.828	45.00	5.551	38.283	38.90	6.959
Planning	14.591	15.00	4.956	10.784	10.80	4.470
Summary Score	87.056	88.30	10.144	74.794	77.50	10.763

For the Patient Safety variable, there were 19 subjects for which the Non-Course faculty ratings scores) were lower than the Course Faculty and 6 subjects with Non-Course faculty having higher scores than Course Faculty (z= -2.811, p=.005). For the Assessment variable, 30 subjects were given lower scores by Non-Course faculty, and 12 subjects received higher ratings

from Non-Course faculty than from Course Faculty (z= -3.277; p = .001). For the Planning variable, 32 subjects were given lower scores by Non-Course faculty than Course Faculty and 10 subjects received higher scores from Non Course faculty (z= -3.227; p = .001). Finally, for Summary Score, 36 subjects were given lower scores by Non-Course faculty raters than Course faculty and 8 subjects who received higher scores from Non-Course Faculty Course Faculty (z= -4.452; p < .001). This comparison of scores between rater groups is presented in Table 4-5.

Table 4-5 Comparison of Scores Between Rater Groups

Variable	*Comparison of Scores by Group*
Patient Safety	19/44 subjects were scored lower by Non-course faculty than by Course faculty
Assessment	30/44 subjects were scored lower by Non-course faculty than by Course faculty
Planning	32/44 subjects were scored lower by Non-course faculty than by Course faculty
Summary Score	36/44 subjects were scored lower by Non-course faculty than by Course faculty

Differences between the rater groups on Medication ratings were examined with a McNemar test. There was no significant agreement between groups of raters on Medication (p=.002). There were significant differences between rater groups on Medication. Course faculty had 4 students who failed and 40 students who passed. Non-course Faculty had 16 students who failed and 28 students who passed. The groups agreed on 3 students who failed and 27 students who passed. See Table 4-6 below.

Table 4-6 Differences Between Rater Groups on Medication

	Non-Course Faculty		
	Fail	Pass	Total
Course Faculty			
Fail	3	1	4
Pass	13	27	40
Total	16	28	44

Research question 4. What were the inter-rater reliability coefficients for each subscale on the revised OSCE instrument when scored by multiple raters?

Agreement between groups of ratings was performed using the revised OSCE instrument. Measurement of the extent to which data collectors or raters assign the same score to

the same variable is called interrater reliability or interrater agreement. Interrater reliability or agreement is also referred to as the level of agreement between multiple raters using the same instrument on the same group of subjects. Agreement is necessary when data are collected by observers or raters as it represents consistency among the ratings of observations. (Hallgren, 2012; Lim, Palethorpe & Rodger, 2012). A Spearman's rho test was computed to analyze agreement or association between the two groups of raters (Course faculty and Non-Course faculty) on the variables of Patient Safety, Assessment, Planning and Medication. There was no significant agreement between rater groups on ratings of student performances (See Table 4-7 below).

Table 4-7 Analysis of Agreement between Course Faculty and Non-Course Faculty

Variables	*Spearman Rho*	*(p)*
Patient Safety	-.103	.506
Assessment	-.236	.123
Planning	-.053	.732
Summary Score	.064	.678

For Summary Score, the variable was recoded to reflect the scores that corresponded to passing scores, which was 85% and failing, scores below 85%. The expectation was that students must achieve 85% on the variables Patient Safety, Assessment, and Planning, and 100% on the Medication variable. Scores were given a binomial pass/fail code to determine whether there was agreement between Course Faculty and Non-Course Faculty on whether a student had passed or failed. It should be noted that the Non-Course Faculty were not involved in making actual pass/fail decisions on student performances. A kappa statistic was computed for this variable in order to measure the level of agreement compared to how much agreement would be expected to be present by chance alone, or measurement error (Lim, Palethorpe, & Rodger, 2012; Viera & Garrett, 2005). There was no significant agreement between Course Faculty and Non-course faculty on whether students had achieved a passing score on Patient Safety, Assessment, and Planning. (See Table 4-8 below).

Table 4-8 Agreement between Course Faculty and Non-Course Faculty

Variables	*Kappa*	*(p)*
Summary Score Pass/Fail	-.007	.941

Analysis of Internal Consistency of Subscales on the revised OSCE Instrument

Internal consistency measures the degree to which items on a scale measure the same construct or behavior (Pallant, 2013). The database contained data only for each subscale of the revised OSCE instrument. The internal consistency of the four subscales on the revised OSCE instrument was examined with a Spearman rank-order correlation coefficient. There was fair consistency between the categories of Assessment and Planning by both Course faculty and Non-Course Faculty. There was no or low consistency among the other subscale scores. Results were presented in Tables 4-9 and 4-10 below for Course Faculty and Non-course Faculty, respectively.

Table 4-9 Analysis of Internal Consistency of Subscales Among Course Faculty

Course Faculty

	Patient Safety	Assessment	Planning	Medication
Patient Safety	1.000	.035 p=.823	.177 p=.251	-.016 p=.918
Assessment	.035 p=.823	1.000	.611** p<.001	-.162 p=.294
Planning	.177 p=.251	.611** p<.001	1.000	-.255 p=.095
Medication	-.016 p=.918	-.162 p=.294	-.255 p=.095	1.000

Table 4-10 Analysis of Internal Consistency of Subscales Among Non-Course Faculty

	Non-course Faculty			
	Patient Safety	Assessment	Planning	Medication
Patient Safety	1.000	-.251 p=.100	-.273 p=.073	-.209 p=.173
Assessment	-.273 p=.073	1.000	.717** p<.001	.212 p=.167
Planning	-.251 p=.100	.717** p<.001	1.000	.039 p=.800
Medication	-.209 p=.173	.212 p=.167	.039 p=.800	1.000

Summary

In this database containing 80 subjects, representing student performances on a senior-level end of course OSCE, 44 students were rated by six Course Faculty and three Non-course Faculty using a revised OSCE instrument. There were significant differences in ratings among Course Faculty in the variables of Patient Safety, Assessment, Planning and Summary Score. There were no significant differences among Non-course Faculty these same variables, and the variable Medication.

There were significant differences between the two rater groups (Course Faculty and Non-course Faculty) on the variables of Patient Safety, Planning, Assessment and Summary Score. Overall course faculty rated subjects with higher scores than did non-course faculty. There was no agreement between the two rater groups on the variables of Patient Safety, Planning, Assessment, and Summary Score, nor for the binomial variables of Pass/Fail on the summary score of the OSCE and the Medication competency. There was fair consistency between the subscales of Assessment and Planning, but there was low or no consistency between the other subscale scores.

Chapter 5

Discussion

Introduction

A retrospective study using secondary data analysis was designed to examine reliability of a revised, faculty-developed OSCE instrument when used by multiple raters. The study methodology was discussed in chapter three.

Results of the data analyses were reported in Chapter 4 including differences among Course Faculty raters, differences among Non-course Faculty raters, and then differences between the two groups of raters. There were significant differences among Course Faculty raters on the variables of Patient Safety, Assessment, Planning, Medication, and a computed variable of Summary Score. There were no significant differences between Non-course Faculty raters on the variables of Patient Safety, Assessment, Planning, Medication and the computed variable of Summary Score. There was no agreement between groups of raters on these same variables. The findings showed fair consistency between the category variables of assessment and planning; with low or no consistency on the other categories.

Interpretation of Findings

Overall Ratings of Student Performance

Nurse educators must evaluate instruments used in high-stakes testing situations to determine reliability, validity, and rater reliability. Evaluation instruments must be designed to yield similar results when used by multiple faculty raters across time and cohorts of students. Newly developed or revised instruments must be tested because of the possibility of random error that may occur in any measurement instrument (Cordi, et al., 2012). When the stakes are high, faculty must be objective in evaluating students without regard to grade inflation (Donaldson & Gray, 2011) or the possibility of failing to fail a student who has not met the performance criteria (Docherty & Dieckmann, 2015).

Research Questions

Research question 1

Were there differences among the six Course Faculty's ratings of student performance using the revised OSCE instrument? Course Faculty raters consistently assigned higher ratings for student performance on Patient Safety, Assessment, Planning, and the Summary Score than did Non-course Faculty raters. Summary Score was the total score of the other categories (Patient Safety worth 30 points, Assessment worth 50 points, and Planning worth 20 points). During the OSCE performances, a summary score of 85% was set as the standard for successful completion of the OSCE. Successful completion of the OSCE, defined as achievement of 85% on the variables of Patient Safety, Assessment, and Planning, and 100% achievement of the Medication category, was required prior to students beginning the final clinical course prior to graduation. Non-course Faculty rated a greater number of student performances as falling short of the 85% score (n=36) than did Course Faculty (n=16)

All of the lowest ranks came from one faculty member, however this faculty performed only 4 ratings of student performance using the new OSCE tool. This faculty member rated 4 subjects using the new tool and 4 subjects using the old tool. The reasons for this are unclear, yet may represent a part—time faculty member who supervised fewer students The highest mean rankings for patient safety were 27.00 (out of 44 rankings). Two faculty raters had a mean rank of 27.00; one faculty rated 4 subjects with the new tool, and all 4 received the maximum score of 30; the second faculty rated 7 subjects using the new tool, and all 7 received the highest possible score of 30. For planning the highest mean ranking was 33.64, which corresponds to a range of scores for 7 subjects of 15-20, with the mean score 19. Finally, the highest mean ranking for summary score was 32.64, which corresponds to a range of scores of 87-100 for 7 subjects, with the mean score 94.85. Three of the highest mean rankings (corresponding to highest scores) came from one faculty member. That faculty member rated a total of 10 students, 7 with the new tool and 3 with the old tool. Again the reasons for the highest and lowest scores coming from predominantly the same two faculty members is unclear, but could represent part-time faculty

members who were less familiar with the elements of the new tool. If faculty raters were unclear about criteria for rating specific elements of the tool, or whether subjects had indeed met the criteria, they may have erred on the side of the student, yielding higher scores, while other faculty may have erred on the side of patient safety, providing lower subject scores, believing that was a more critical element.

There was no significant difference among Course Faculty raters for the variable Assessment. The total possible score for Assessment was 50, and the mean score from all faculty raters on this variable was 43.83. The elements included in this variable included correct performance of vital signs, correctly performing a focused assessment of the primary body system, including assessment of the intravenous (IV) site, and formation of a nursing diagnosis. Nursing faculty may have found these elements more objective in nature, and more easily quantified, yielding more consistent scores. For the Medication variable, the sample size was too small to elicit meaningful results.

Because most studies in the literature did not address faculty rating students on end-of-course, high-stakes performances during an OSCE, it is difficult to compare the results of this study to previous studies. Previous studies included pilot studies designed to develop an OSCE tool (Walsh, Bailey, et al., 2010); studies in which the student rated satisfaction with the OSCE (Rentschler, et al., 2010); voluntary participation by students (Merrimayn, Stayt & Ricketts, 2014; Oranye, Ahmad, C., Ahmad, A., & Bakar, 2012; Raurell-Torreda, et al., 2015); participants who were practicing nurses, enrolled in an open distance learning course (Oranye, et al., 2012); and studies in which the OSCE checklist was developed as a formative evaluation and used to compare a simulation group with a traditional group (Raurell-Torreda, et al., 2015; Merrimay, Stayt & Ricketts, 2014). Often these studies did not include the OSCE checklist tool nor did they report reliability or validity results.

Observations of performance are subject to perception or human judgment, where different raters may perceive performance differently from one another. Rater training is a method employed to provide optimal results when multiple raters are involved. Raters must have a clear

understanding of the instruments, the content, and purpose of evaluation (Adamson & Kardong-Edgren, 2012); may be asked to review expected behaviors and/or asked to view a sample video showing various levels of performance (Rizzolo, Kardong-Edgren, Oermann, & Jeffries, 2015; and may be asked to meet to conduct practice scoring sessions prior to implementation of a new tool (Moskal & Leydens, 2000). The INACSL Standard VII, criterion 2 concerning summative evaluation states that summative evaluation should be explained before the start of the evaluation process and conducted by trained objective observers or raters (Sando, et al., 2013). The lack of rater training prior to implementation of this new tool could explain some of the differences in consistency of ratings between course faculty raters.

Research question 2

Were there differences among the three Non-course Faculty ratings of student performance using the revised OSCE instrument?

There were no significant differences among the three Non-course faculty raters on each of the study variables of Patient Safety, Assessment, Planning, Medication, and the computed variable of Summary Score. In many of the published studies, raters were blinded, either by having no relationship with the student or in randomized, controlled trials, where students were randomly assigned to groups (Merriman & Stayt, 2014; Hayden, Keegan, Kardong-Edgren & Smiley, 2014a). The INACSL Simulation Best Practices Standard VII, criterion 3 regarding high-stakes evaluation states that evaluation of participants' performance by objective observers or raters increases objectivity and diminishes biased assessment (Sando, et al., 2013). The ratings from these non-course faculty may have been more consistent because they were not directly involved in teaching or evaluation of the students, thus adding a greater degree of objectivity.

Research question 3

Were there differences between ratings of Course Faculty and Non-course Faculty on student performance using the revised OSCE instrument?

A significant finding from this study was the differences between rater groups in assigning ratings of student performance on the variables of Patient Safety, Assessment, Planning and

Summary Score. Course faculty consistently assigned higher (better) scores than did External faculty raters on these variables. For Summary Score, the Non-Course Faculty rated 36 of the 44 subjects with lower (poorer) scores than did Course Faculty.

Grade inflation and the concept of “failure to fail” could play a role in course faculty assigning higher grades on evaluations. Donaldson & Gray (2012) cite possible reasons for grade inflation, defined as a greater percentage of higher scores than may be a true reflection of actual performance. They reported that grade inflation could result from student expectations and pressure on faculty, faculty reluctance to fail students, evaluator-student relationships, and tool design. When discussing tool design, Donaldson & Gray (2012) explained that equal weighting of objectives could lead to grade inflation when students could succeed overall, yet miss the “big picture”. Donaldson & Gray (2012) further discuss how during face to face evaluations, faculty may be reluctant to give negative feedback during evaluation of student performance. They noted that this was not limited to inexperienced evaluators, but even experienced assessors, especially those who were non-tenured and relied more heavily on positive student evaluations for their own positive evaluation and promotion were likely to demonstrate grade leniency. Faculty who have a student-evaluator relationship or bond with students may become subject to what is known as the “halo effect”, where a rater’s evaluation of a student’s performance may be influenced by his or her overall impression of the student in other traits (Iramaneerat & Yudkowsky, 2007). Further, faculty may be reluctant to evaluate the student as having failed the performance because of the excess work that is involved with remediation and retesting (Danyluk, et al., 2015). Little research has been done in nursing education regarding using blinded reviewers to evaluate student performances, focusing instead on reviewers being blinded to teaching methods (Gundrosen, Soligard, & Aadahl (2014). The literature is clear in recommending that faculty mentors not be assigned to evaluate their own students (Donaldson & Gray, 2012; Isaacson & Stacy, 2009). Faculty who design clinical courses may consider using trained, non-course faculty to evaluate student performances on high-stakes examinations.

Strategies to improve rater consistency include having faculty discuss in advance, each competency or clinical behavior, what it means, and what to look for when evaluating the competency, as well as standards for determining a passing or failing grade (Rizzollo, Kardong-Edgren, Oermann & Jeffries, 2015; Oermann, Yarbrough, et. al., (2009a).

Research question 4

What were the inter-rater reliability coefficients for each subscale on the revised OSCE instrument when scored by multiple raters?

An important finding from this study was that using the revised OSCE instrument, there was no agreement between rater groups (Course Faculty and Non-course Faculty) on ratings of Patient Safety, Assessment, Planning, and Summary Score. The Spearman correlation coefficients were negative for Patient Safety, Assessment and Planning, and showed a small or weak correlation. Correlations of 0.10 to 0.29 are considered small (Cohen, 1988).

Additionally there was no agreement between rater groups on Summary Score of the OSCE performances that were considered passing (85% or greater). The kappa statistic is intended to provide a quantitative measure of the magnitude of agreement between observers (Viera & Garrett, 2005). Values <0 indicate less than chance agreement and kappa 0.01 – 0.20 indicating slight or fair agreement. There was fair agreement between rater groups on the Medication variable, determining whether students had passed or failed this element.

Internal Consistency of the Revised OSCE Tool

Internal consistency for the subscales on the revised OSCE tool demonstrated fair internal consistency between the subscales of Assessment and Planning. The other subscales demonstrated no or low consistency.

Recent studies evaluating reliability and validity of simulation evaluation instruments have included results for internal consistency. The Creighton Simulation Evaluation Instrument was evaluated and Cronbach's alpha reported as .979 (Adamson, et al., 2011; Adamson & Kardong-Edgren, 2012). In Adamson and Kardong-Edgren (2012), reliability and validity of two additional instruments was reported. The Lasater Clinical Judgment Rubric yielded Cronbach's alpha of

.974 and the Seattle University Evaluation Tool yielded Cronbach's alpha of .965 (Adamson, Gubrud, Sideras & Lasater, 2011).

Additional refinement and testing of the revised OSCE instrument would be warranted. A Cronbach's alpha or Spearman's rho coefficient of 1.00 indicates perfect reliability. Typically a reliability of 0.80 is an indication of strong consistency, and for new instruments a reliability of 1.70 is considered acceptable (Grove, Gray & Burns, 2015).

Limitations

A retrospective, secondary data analysis limited the number of variables and data that could be collected. Using data previously collected resulted in the same limitations found in the prior study.

Conclusions

Nurse educators need reliable and valid tools to evaluate student performance, yet few tools have been developed or tested for psychometric properties. This secondary analysis was conducted to evaluate interrater reliability of the revised OSCE instruments. The results of this study indicated that the revised OSCE instrument did not demonstrate sufficient rater reliability for use in high-stakes clinical evaluation. Course faculty raters may have been reluctant to allow students to fail the performance, and thus they assigned higher scores than were warranted per the criteria on the OSCE instrument. Faculty evaluators also may find it difficult to allow students to fail during face to face evaluations, compared to other methods of evaluation. The lack of rater training may have contributed to faculty giving students the benefit of the doubt, if they were uncertain about how items on the instrument should be scored. For high-stakes performance evaluation using faculty raters who are objective, and not associated with the course may provide more accurate evaluations of student performance. Faculty must continue to conduct rigorous research to refine and test OSCE instruments.

Implications for Nursing

Nurse educators need to use reliable and valid instruments to evaluate students in high-stakes performance evaluations. This OSCE instrument will need additional refinement and testing to improve reliability.

Educators may elect to implement and test instruments in their educational settings that have previously been shown to demonstrate reliability and validity, such as the Creighton Competency Evaluation Instrument (Rizzolo, et al., 2015; Todd, et al., 2008) and the Lasater Clinical Judgment Rubric (Lasater, 2008).

In 2013, the International Nursing Association for Clinical Simulation and Learning developed Standards of Best Practice for simulation, assessment and evaluation (Sando, et al., 2013). The standards include criterion for formative assessments, summative evaluation and high-stakes evaluation. Nurse educators can implement these guidelines in their colleges of nursing for assessing students in simulation or individual performance to standardize evaluation practices. Implementation of these guidelines may allow faculty to avoid inconsistencies involved with all aspects of evaluation and may eliminate the risk for grade inflation and failure to fail students.

Recommendations for Future Studies

The NLN Evaluation of Learning Advisory Council study (Oermann, et al., 2009a; Oermann, et al., 2009b) surveyed faculty in nursing programs to determine the strategies and methods used to evaluate students in the affective domain of learning. Seventy-seven respondents indicated they used OSCEs, however the literature rarely includes examples of actual OSCE tools used. An initial study would be conducted to survey faculty regarding the items included on their OSCEs and rationale for inclusion, as well as to share their tools for comparison purposes. Faculty are likely using many of the same elements for evaluation. This would affirm the inclusion or exclusion of nursing actions that are being included as part of high-stakes evaluations.

Nurse educators may wish to continue to use this OSCE instrument following additional refinement and testing to establish rater reliability. Following the method established by Adamson & Kardong-Edgren (2012), a larger sample of raters should be recruited. Rater training should be provided to include viewing videos of leveled performances and participation in meetings to discuss questions about instrument items, for optimal consistency in use.

In the INACSL Standards of Best Practice (Sando, et al., 2013), one of the guidelines addresses the use of trained objective observers or raters. Little research has been conducted specific to using blinded reviewers to evaluate student performance in lieu of ratings from course faculty. A pilot study could be conducted to recruit and train faculty members with expertise in simulation and objective structured clinical examinations (OSCEs), and not associated with the course, to provide evaluation of student performances in low stakes or formative situations. Results from that study could be used to guide future research for using blinded reviewers in high stakes situations.

The Creighton Competency Evaluation Instrument has demonstrated validity and rater reliability in high-stakes evaluation. Nurse educators would be wise to replicate those studies for use in their programs, to increase the database of information and to determine whether this tool could replace faculty-developed tools for increased consistency of evaluation.

Summary

Few studies have been conducted to evaluate instruments for evaluation of individual student performance during high stakes situations or OSCEs. In this study, the revised OSCE checklist tool did not demonstrate that it was reliable when used by multiple raters. Although the results could be viewed as disappointing, educators need sound research to determine whether or not their methods are effective. It is important for nurse educators to learn that a tool may not be the best tool for evaluation purposes, so that either continued tool refinement and research can occur, or alternate tools may be selected.

The findings of this study contribute to the knowledge base for nurse educators to consider when designing evaluation strategies for students in clinical performances. The findings

of this study can be used by the Course faculty to make decisions regarding best practices for evaluation of students in high-stakes performances on an OSCE.

Appendix A

Data Collection Tool

Example of new instrument (specific criteria not included)

Evaluation Criteria Assign 1 point for "Done", 0 points for "Not Done"	Done	Not Done	Points
Patient Safety 5 points			
1. Identifies self to patient *(at least once)*			
2. Identifies patient:			
a. Name			
b. Date of birth			
3. Performs appropriate hand hygiene			
4. Safety measures			
Assessment & Nursing Diagnosis 13 points Initial VS for patient scenario			
5. Vital signs:			
a. Takes BP correctly			
b. Takes pulse correctly			
c. Counts respiratory rate correctly			
6. Focused assessment			
a. Assessment:			
• Identifies abnormal findings			
• Performs auscultation			
• Identifies patient breath sounds			
• Identifies patient symptoms			
b. Cardiac Assessment:			
• Auscultates			
• Identifies signs and symptoms			
• Identifies findings			
c. Recognizes patient complaints			
d. IV assessment			
e. Student recognizes need for intervention			
Interventions and Evaluation/Re-assessment (Critical Thinking & Decision-Making) 10 points			
7. Follows provider orders for correct intervention			
Evaluation/Re-assessment after intervention			
8. Identifies abnormal findings			
9. Auscultates			
10. Identifies lung sounds			
11. Checks provider orders			
12. Implements correct interventions			
Evaluation/Re-assessment after intervention			
13. Identifies correct findings			
14. Auscultates			
15. Identifies changes			
16. Notes VS changes			

Medication Administration	10 points		
17. Identifies right patient			
18. Assesses patient for allergies			
19. Administers right drug			
20. Cleans top of vial			
21. Administers right dose			
22. Cleans IV port			
23. Uses correct injection port			
24. Right time			
25. Right rate			
26. Maintains sterility			
Student points (out of 38 total):			

References

Abbott, K.V., Barton, F.B., Terhorst, L. & Shembel, A. (2016). Retrospective studies: A fresh look. *American Journal of Speech-Language Pathology, 25,*157-163.

Adamson, K.A., Gubrud, P., Sideras, S. & Lasater, K. (2011). Assessing the reliability, validity and use of The Lasater Clinical Judgment Rubric: Three approaches. *Journal of Nursing Education, 51*(2), 66-73.

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

Adamson, K.A., Kardong-Edgren, S., & Willhaus, J. (2013). An updated review of published simulation evaluation instruments. *Clinical Simulation in Nursing, 9, e393-e400.*

Adamson, K.A., Parsons, M.E., Hawkins, K., Manz, J.A., Todd, M. and Hercinger, M. (2011). Reliability and internal consistency findings from the C-SEI. *Journal of Nursing Education, 50*(10), 583-586.

Adamson, K. A. & Prion, S. (2012). Making sense of methods and measurement: reliability. *Clinical Simulation in Nursing, 8*(6), e259-e260. http://dx.doi.org/10.1016/j.ecns.2012.05.003.

Aiken, L.H., Clarke, S.P., Cheung, R.B., Sloan, D.M., Silber, J.H. (2003). Educational levels of hospital nurses and surgical patient mortality. Journal of the American Medical Association, 290(12), 1617-1623.

Aiken L, Clarke, S., Sloane D, Sochalski, & J, Silber J. (2002). Hospital nurse staffing and patient mortality, nurse burnout, and job dissatisfaction. *Journal of the American Medical Association* 288, 1987–1993.

American Association of Colleges of Nursing (2008). *Essentials of Baccalaureate Education for Professional Nursing Practice.* Retrieved from http://www.aacn.nche.edu/education-resources/BaccEssentials08.pdf

American Nurses Association, (2015). *Nursing: Scope and standards of practice*, (3nd edition). Washington, D.C.: American Nurses Publishing.

Bartfay, W., Romborough, R., Howse, E., and Leblanc, R. (2004). The OSCE approach in nursing education. *Canadian Nurse 100*(3), 18-23.

Beaumont, K., Luettell, D, & Thomson, R. (2008). Deterioration in hospital patients: early signs and appropriate actions. *Nursing Standard, 23*(1), 43-48.

Benner, P., Tanner, C., & Chesla, C. (2009). *Expertise in nursing practice: caring, clinical judgment, and ethics.* New York: Springer.

Benner, P., Sutphen, M., Leonard, V., & Day, L. (2010). *Educating Nurses: a Call for Radical Transformation,* Jossey-Bass, San Francisco, CA.

Benner, P. (1984). *From novice to expert.* Menlo-Park, California: Addison-Wesley Publishing Company.

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

Black, S., Curzio, J. & Terry, L. (2014). Failing a student nurse: a new horizon of moral courage. *Nursing Ethics, 21*(2), 224-238.

Bogossian, F., Cooper, S., Cant, R., Beauchamp, A., Porter, J, Kain, V., Bucknall, T., Phillips, N.M., and the FIRST2ACT Research Team (2014). Undergraduate nursing students' performance in recognizing and responding to sudden patient deterioration in high psychological fidelity simulated environments: an Australian multi-centre study. *Nurse Education Today, 34,* 691-696.

Buykx, P., Cooper, S., Kinsman, L., Endacott, R., Scholes, J., McConnell-Henry, T., and Cant, R. (2012). Patient deterioration simulation experiences: Impact on teaching and learning. *Collegian, 19,* 125-129.

Buykx, P., Kinsman, L., Cooper, S., McConnell-Henry, T., Cant, R., Endacott, R., & Scholes, J. (2011). FIRST2ACT: Educating nurses to identify patient deterioration – a theory-based model for best practice simulation education. *Nurse Education Today, 31,* 687-693.

Cant, R.P. & Cooper, S.J. (2014). Simulation in the internet age: The place of web-based simulation in nursing education. An integrated review. *Nurse Education today, 34,* 1435-1442. http://dx.doi.org/10.1016/j.nedt.2014.08.001

Cant, R.P. & Cooper, S.J. (2010). Simulation-based learning in nurse education: Systematic review. *Journal of Advanced Nursing, 66*, 3-15.

Cant, R., McKenna, L., & Cooper, S. (2013). Assessing preregistration nursing students' clinical competence: A systematic review of objective measures. *International Journal of Nursing Practice, 19*, 163-176.

Cazzell, M. and Howe, C. (2012). Using objective structured clinical evaluation for simulation evaluation: checklist considerations for inter-rater reliability. *Clinical Simulation in Nursing, 8,* e219-e225.

Chee, J. (2013). Clinical simulation using deliberate practice in nursing education: A Wilsonian concept analysis. *Nurse Education in Practice, 14,* 247-252.

Cicero, T. & Mikasa, A. (2008). Seattle University Evaluation Tool. Contact the developers at CICERO@seattleu.edu or mikasaa@seattleu.edu.

Christensen, M. (2010). Advancing nursing practice: redefining the theoretical and practical integration of knowledge. *Journal of Clinical Nursing, 20,* 873-881.

Cioffi, J. (2000). Recognition of patients who require emergency assistance: A descriptive study. *Heart & Lung, 29*, 262-268.

Clarke, S.P. & Aiken, L.H. (2003). Failure to rescue. *American Journal of Nursing, 103*(1), 42-47.Collins, T., Price, A., & Angrave, P. (2006). Pre-registration education; making a difference to critical care/ *Nursing in Critical Care, 11*(1), 52-57.

Cohen, J.W. (1988). Statistical power analysis for the behavioral sciences (2nd edition). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cooper, S., Cant, R., Bogossian, F., Kinsman, L., Bucknall, T., and the FIRST2ACT Research Team. (2015). Patient deterioration education: Evaluation of face-to-face simulation and e-simulation approaches. *Clinical Simulation in Nursing, 11,* 97-105.

Cooper, S., Cant, R., Porter, J., Missen, K., Sparkes, L., McConnell-Henry, T., and Endacott, R. (2013). Managing patient deterioration: assessing teamwork and individual performance. *Emergency Medicine Journal, 30,* 377-381. doi:10.1136/emermed-2012-201312.

Cooper, S., McConnell-Henry, T., Cant, R., Porter, J., Missen, K., Kinsman, L., Scholes, J., & Endacott, R. (2011). Managing deteriorating patients: Registered nurses' performance in a simulated setting. *Open Nursing Journal, 5,* 120-126.

Cooper, S., Kinsman, L, Buykx, P., McConnell-Henry, T., Endacott, R. & Scholes, J. (2010). Managing the deteriorating patient in a simulated environment: nursing students' knowledge, skills and situation awareness. *Journal of Clinical Nursing, 19,* 2309-2318.

Cordi, V.L., Leighton, K., Ryan-Wenger, N., Doyle, T.J., and Ravert, P. (2012). History and development of the simulation effectiveness tool (SET). *Clinical Simulations in Nursing, 8,* 199-210.

Danyluk, P.J., Luhanga, F., Gwekwerere, Y.N., MacEwan, L., & Larocque, S. (2015). Failure to fail in a final pre-service teaching practicum. *The Canadian Journal for the Scholarship of Teaching and Learning, 6*(3), 1-14.

Dieckmann, P., Gaba, D., & Rall, M. (2007). Deepening the theoretical foundations of patient simulation as social practice. *Simulation in Healthcare, 2*, 183-193.

Dochterman, J., Titler, M., Wang, J., Reed, D., Pettit, D., Matthew-Wilson, M., Budreau, G., Bulecheck, G., Kraus, V., and Kanak, M. (2005). Describing use of nursing interventions for three groups of patients. *Journal of Nursing Scholarship, 37*(1), 57-66.

Dochterman, J.M. & Bulechek, G.M. (Eds.) (2004). *Nursing Interventions Classifications (NIC),* (4th. ed.). St. Louis, MO: Mosby.

Docherty, A. & Dieckmann, (2015). Is there evidence of failing to fail in our schools of nursing? *Nursing Education Perspectives, 36*(4), 226-231.

Donaldson, J.H. & Gray, M. (2011). Systematic review of grading practice: Is there evidence of grade inflation? *Nurse Education in Practice, 12*, 101-114.

Doolan, D.M. & Froelicher, E.S. (2009). Using an existing data set to answer new research questions: a methodological review. *Research and Theory for Nursing Practice: An International Journal, 23*(3), 203-215.

Doolen, J. (2015). Psychometric properties of the Simulation Thinking Rubric to measure higher order thinking in undergraduate nursing students. *Clinical Simulation in Nursing, 11,* 35-43.

Dougherty, C.M. (1999). Surveillance. In: G.M. Bulecheck & J.C. McCloskey (Eds.) Nursing interventions: Effective nursing treatments (3rd ed., pp. 524-533). Philadelphia: Saunders.

Douw, G., Schoonhoven, L., Holwerda, T., Huisman-de-Waal, G., van Zanten, A.R., van Achterberg, T., & van der Hoeven, J.G. (2015). Nurses' worry or concern and early recognition of deteriorating patients on general wards in acute care hospitals: a systematic review. *Critical Care, 19*, 230 (1-11).

Dreyfus, S.E. & Dreyfus, H.L. (1980). *A five-stage model of the mental activities involved in directed skill acquisition.* Berkeley, CA: University of California, Operations Research Center. Retrieved from www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA084551

Dreyfus, S.E. & Dreyfus, H.L. (1986). *Mind over machine: The power of human intuition and expertise in the era of the computer*. New York: Collier MacMillan.

Eldair, S., Sebaae, H., Feky, H., Hussein, H., Abd El Fadil, N., and El Shaer, I.H. (2010). An introduction of OSCE versus traditional method in nursing education: faculty capacity building and students' perspectives. *Journal of American Science, 22,* 60-67.

Endacott, R., Scholes, J., Cooper, S., McConnell-Henry, T., Porter, J., Missen, K., Kinsman, L., and Champion, R. (2011). Identifying patient deterioration: using simulation and reflective interviewing to examine decision making skills in a rural hospital. *International Journal of Nursing Studies, 49,* 710-717.

Endacott, R., Scholes, J., Buykx, P., Cooper, S., Kinsman, L., and McConnell-Henry, T. (2010). Final-year nursing students' ability to assess, detect and act on clinical cues of deterioration in a simulated environment. *Journal of Advanced Nursing, 66*(12), 2722-2731. doi:10.1111/j.1365-2648.2010.05417.x.

Ericsson, K.A., Krampe, R.T., and Tesch-Romer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*(3), 363-405.

Ericcson, K.A., Whyte, J. & Ward, P. (2007). Expert performance in nursing: Reviewing research On expertise in nursing within the framework of the expert-performance approach. *Advances in Nursing Science, 30*(1), E58-E71.

Fasolino, T. & Verdin, T. (2015). Nursing surveillance and physiological signs of deterioration. *MedSurg Nursing 24*(6), 397-402.

Fero, L.J., O'Donnell, J.M., Zullo, T.G., Dabbs, A.D., Kitutu, J., Samosky, J.T. & Hofman, L.A. (2010). Critical thinking skills in nursing students; comparison of simulation-based performance with metrics. *Journal of Advanced Nursing, 66*(10), 2182-2193.

Fisher, D. & King, L. (2013). An integrative literature review on preparing nursing students through simulation to recognize and respond to the deteriorating patient. *Journal of Advanced Nursing, 69*(11), 2375-2388.

Gallant, M., MacDonald, J. A., & Smith Higuchi, K. A. (2006).A remediation process for nursing students at risk of clinical failure. *Nurse Educator, 31*,223-227.

Gates, M., Parr, M.B. & Hughen, J. (2012). Enhancing nursing knowledge using high-fidelity simulation. *Journal of Nursing education, 51*(1), 9-15.

Gordon, C.J. & Buckley, T. (2009). The effect of high-fidelity simulation training on medical-surgical graduate nurses' perceived ability to respond to patient clinical emergencies. *The Journal of Continuing Education in Nursing, 40*(11), 491-498.

Grove, S., Gray, J.R. and Burns, N. (2015). *Understanding Nursing Research: Building an Evidence-Based Practice.* St. Louis, Missouri: Elsevier.

Gundrosen, S., Solligard, E., & Aadahl, P. (2014). Team competence among nurses in an intensive care unit: The feasibility of *in situ* simulation and assessing non-technical skills. *Intensive and Critical Care Nursing, 30*, 312-317. http://dx.oi.org/10.1016/j.iccn.2014.06.007

Hall, S.W. (2015). High-Fidelity Simulation for Senior Maternity Nursing Students, *Nursing Education Perspectives, 36*(2), 124-126.

Hallgren, K.A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology 8*(1), 23-34.

Harden, R., Stevenson, M., Wilson Downie, W., and Wilson, G. (1975). Assessment of clinical competence using objective structured clinical examination. *British Medical Journal, 1*(5955), 447-451.

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

Hayden, J.K., Smiley, R.A., Alexander, M., Kardong-Edgren, S. and Jeffries, P.R. (2014b). The NCSBN National Simulation Study: A longitudinal, randomized, controlled study replacing clinical hours with simulation in prelicensure nursing education. *Journal of Nursing Regulation, July 2014 Supplement, 5*(2), S1-S64.

Henneman, E.A., Cunningham, H., Roche, J., & Curnin, M. (2007). Human patient simulation: Teaching students to provide safe care. *Nurse Educator, 32,* 212-217.

Henneman, E.A., Gawlinski, A. & Giuliano, K.K. (2012). Surveillance: A strategy for improving patient safety in acute and critical care units. *Critical Care Nursing 32*(2), e9-e18.

INACSL (2011). Standards of best practice: Simulation [Special Issue]. *Clinical Simulation in Nursing 7*(4S), S18-S19.

Iramaneerat, C. & Yudkowsky, R. (2016). Rater errors in a clinical skills assessment of medical students. *Evaluation & the Health Professions, 30*(3), 266-283.

Isaacson, J.J. & Stacy, A.S. (2009). Rubrics or clinical evaluation: Objectifying the subjective experience. *Nurse Education in Practice, 9*, 134-140.

Jarvis, C. (2012). *Physical Examination and Health Assessment, (6th. Ed.)*. St. Louis, Missouri: Elsevier Saunders.

Jeffries, P.R. & Norton, B. (2005). Selecting learning experiences to achieve curriculum outcomes. In D.M. Billings and J.A. Halstead (Eds.). Teaching in Nursing: A guide for faculty (2nd ed., pp. 187-212). St. Louis, MO: Elsevier.

Jones, A., Pegram, A., and Fordham-Clarke, C. (2010). Developing and examining an Objective Structured Clinical Examination. *Nurse Education Today, 30,* 137-141.

Kardong-Edgren, S., Adamson, K.A. & Fitzgerald, C. (2010). A review of currently published evaluation instruments for human patient simulation. *Clinical Simulations in Nursing, 6, 25-35.*

Kelly, L. & Vincent, D. (2011). The dimensions of nursing surveillance: a concept analysis. *Journal of Advanced Nursing, 67*(3), 652-661.

Kutney-Lee, A., Lake, E.T., & Aiken, L.H. (2009). Development of the hospital nurse surveillance capacity profile. *Research in Nursing and Health, 32, 217-228.*

Lamont, S. & Brunero, S. (2013). 'e-Simulation' Part 1: Development of an interactive multimedia mental health education program for generalist nurses. *Collegian, 20,* 239-247. http://dx.doi.org/10.1016/j.colegn.2012.11.001

Lamont, S. & Brunero, S. (2014). 'e-Simulation' Part 2: Evaluation of an interactive multimedia mental health education program for generalist nurses. *Collegian, 21*, 3-9 http://dx.doi.org/10.1016/j.colegn.2012.11.002.

Lancia, L., Petrucci, C., Giorgi, F. Dante, A., & Cifone, M.G. (2013). Academic success or failure in nursing students: Results of a retrospective observational study. *Nurse Education Today, 33,* 1501-1505.

Lasater, K. (2007). High-fidelity simulation and the development of clinical judgment: students' experiences. *Journal of Nursing Education, 46*(6), 269-276.

Lauder, W., Holland, K., Roxburgh, M., Topping, K., Watson, R., Johnson, M., Porter, M., & Behr, A. (2008). Measuring competence, self-reported competence and self-efficacy in preregistration nursing students. *Nursing Standard, 22,* 35-43.

Law, M. (2005). Reduce, reuse, recycle: Issues in the secondary use of research data. *IASSIST Quarterly 29.* Retrieved from http://www.iassistdata.org/sites/default/files/iqvol291law.pdf.

Liaw, S.Y., Rethans, J., Scherpbier, A., and Piyanee, K. (2011). Rescuing a patient in deteriorating situations (RAPIDS): A simulation-based educational program on recognizing, responding and reporting of physiological signs of deterioration. *Resuscitation, 82,* 1224-1230.

Lim, S.M., Palethorpe, N., & Rodger, S. (2012). Understanding the common inter-rater reliability measures. *International Journal of Therapy and Rehabilitation, 19*(9), 488-496.

Mark, B.A. & Harless, D.W. (2010). Nurse staffing and post-surgical complications using the present on admission indicator. *Research in Nursing and Health*, 33(1), 35-47.

McCloskey J.C. & Bulechek G.M. (Eds.). (1996). Nursing interventions classification. (2nd ed.) St. Louis, MO: Mosby.

McKenna, L., Missen, K., Cooper, S., Bogossian, F., Bucknall, T., and Cant, R. (2014). Situation awareness in undergraduate nursing students managing simulated patient deterioration. *Nurse Education Today, 34,* e27-e31.

McWilliam, P. and Botwinski, C. (2010). Developing a successful nursing objective structured clinical examination. *Journal of Nursing Education, 49*(1), 36-41. doi:10.3928/01484834-20090915-01.

Merriman, C.D., Stayt, L.C., and Ricketts, B. (2014). Comparing the effectiveness of clinical simulation versus didactic methods to teach undergraduate adult nursing students to recognize and assess the deteriorating patient. *Clinical Simulations in Nursing, 10,* e119-e127.

Mikasa, A.W., Cicero, T.F., and Adamson, K.A. (2013). Outcome-based evaluation tool to evaluate student performance in high-fidelity simulation. *Clinical Simulation in Nursing, 9,* e361-e367.

Minick, P. & Harvey, S. (2003). The early recognition of patient problems among medical-surgical nurses. *MedSurg Nursing, 12*(5), 291-297.

Mitchell, M.L., Henderson, A., Groves, M., Dalton, M., and Nulty, D. (2009). The objective structured examination: (OSCE): Optimising its value in the undergraduate nursing curriculum. *Nurse Education Today, 29,* 398-404.

Moskal, B.M. & Leydens, J.A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research and Evaluation 7*(10). Retrieved from http://PAREonline.net/getvn.asp?v=7&n=10

National Patient Safety Agency (NPSA) (2007a). Recognising and responding appropriately to early signs of deterioration in hospitalized patients. NPSA, London, Retrieved from https://www.nice.org.uk/guidance/cg50/evidence/full-guideline-195219037

National Patient Safety Agency (NPSA) (2007b). Safer care for the acutely ill patient: learning from serious incidents. NPSA, London. Retrieved from http://www.nrls.npsa.nhs.uk/resources/?entryid45=59828

Odell, M., Victor, C., & Oliver, D. (2009). Nurses' role in detecting deterioration in ward patients: systematic literature review. *Journal of Advanced Nursing, 65*(10), 1992-2006.

Oermann, M.H., Saewert, K.J., Charasika, M. & Yarbrough, S.S. (2009a). Assessment and grading Practices in schools of nursing: National survey findings part I. *Nursing Education Perspectives, 30*(5), 274-278.

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

Oermann, M.H., Kardong-Edgren, S., Odom-Maryon, T., Hallmark, B.F., Hurd, D., Rogers, N., Haus, C., J., Keegan-McColgan, Snelson, C., Dowdy, S.W., Resurreccion, L.A., Kuerschner, D.R., Lamar, J., Tennant, M.N., and Smart, D.A. (2011). Deliberate practice of motor skills in nursing education: CPR as exemplar. *Nursing Education Perspectives, 32*(5), 311-315.

Oranye, N.O., Ahmad, C., Ahmad, N., & Bakar, R.A. (2012). Assessing nursing clinical skills competence through objective structured clinical examination (OSCE) for open distance learning students in Open University *Malaysia*. *Contemporary Nurse, 41*(2), 233-241.

Pallant, J. (2013). *SPSS survival manual: a step-by-step guide to data analysis using IBM SPSS (5th ed.).*New York, New York: McGraw-Hill.

Parsons, M.E., Hawkins, K.S., Hercinger, M., Todd, M., Manz, J.A., and Fang, X. (2012).Improvement in scoring consistency for the Creighton Simulation Evaluation Instrument. *Clinical Simulation in Nursing, 8, e233-e238.*

Paskausky, A.L. & Simonelli, M.C. (2014). Measuring grade inflation: A clinical grade discrepancy score. *Nurse Education in Practice, 14*, 374-379. Retrieved from http://dx.doi.org/10.1016/j.nepr.2014.01.011

Preston-Safarz, P. & Bolick, B.N. (2015). A pilot study to implement and evaluate the use of objective structured clinical examinations in an RN to BSN nursing program. *Clinical Simulations in Nursing, 11,* 59-63.

Purling, A. & King, A. (2012). A literature review: graduate nurses' preparedness for recognizing and responding to the deteriorating patient. *Journal of Clinical Nursing, 21,* 3451-3465.

Raurell-Torreda, M., Olivet-Pujol, J., Romero-Collado, A., Malagon-Aguilera, M., Patino-Maso, J. & Baltasar-Bague, A. (2015). Case-based learning and simulation: useful tools to enhance nurses' education? Nonrandomized controlled trial. *Image: Journal of Nursing Scholarship, 47*(1), 34-42.

Rentschler, D.D., Eaton, J., Cappiello, J., McNally, S.F. and McWilliam, P. (2007). Evaluation of undergraduate students using objective structured clinical evaluation. *Journal of Nursing Education, 46*(3), 135-139.

Rich, K. (1999). Inhospital cardiac arrest: Pre-event variables and nursing response. *Clinical Nurse Specialist, 13*(3), 147-153.

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.

Rizzolo, M.A., Oermann, M.H., Jeffries, P. & Kardong-Edgren, S. (2011). NLN project to explore use of simulation for high-stakes assessment. *Clinical Simulation in Nursing,7*(6), e261-262.

Rushforth, H.E. (2007). Objective structured clinical examination (OSCE): review of literature and implications for nursing education. *Nurse Education Today, 27,* 481-490.

Ryan-Wenger, N.A. (2010). Evaluation of measurement precision, accuracy, and error in biophysical data for clinical research and practice. In: *Measurement in Nursing and Health Research, (4th ed.; Waltz, Strickland, and Lenz, eds.).* New York: Springer Publishing Company.

Sando, C.R., Meakim, C., Gloe, D., Decker, S., & Borum, J.C. (2013). Standards of best practice: Simulation Standard VII: Participant assessment and evaluation. *Clinical Simulation in Nursing 9*, e30-e32. http://dx.doi.org/10.1016/j.ecns.2013.04.007.

Schoneman, D. (2002). Surveillance as a nursing intervention: use in community nursing centers. *Journal of Community Health Nursing, 19*(1), 33-47.

Schubert, C.R. (2012). Effect of simulation on nursing knowledge and critical thinking in failure to rescue events. *The Journal of Continuing Education in Nursing, 43*(10), 467-471.

Schuwirth, L. and van der Vleuten, C. (2003). The use of clinical simulations in assessment. *Medical Education, 37* (Supplement), 65-71.

Selim, A.A., Ramadan, F,A., El-Gueneidy, M.M., & Gaaafer, M.M. (2012). Using objective structured clinical examination (OSCE) in undergraduate psychiatric nursing education: Is it reliable and valid? *Nurse Education Today, 32,* 283-288.

Shever, L.L. (2011). The impact of nursing surveillance on failure to rescue. *Research in theory Nursing Practice, 25*(2), 107-126.

Stiller, J.J., Nelson, K.A., Anderson, M., Ashe, M.J., Johnson, S.T., Sandhu, K., Mangold, E., Scheid, S., and LeFlore, J. (2015). Development of a valid and reliable evaluation instrument for undergraduate nursing students during simulation. *Journal of Nursing Education and Practice, 5*(7), 1-8.

Tabassum, N., Saeed, T., Dias, J.M., & Allana, S. (2016). Strategies to eliminate medication error among undergraduate nursing students. *International Journal of Nursing Education, 8*(1), 167-171.

Tanicala, M.L., Scheffer, B.K. & Roberts, M. S. (2011). Pass/fail nursing student clinical behaviors Phase I: moving toward a culture of safety. *Nursing Education Perspectives, 32*(3), 155-161.

Tanner, C. (2006). Thinking like a nurse: A research-based model of clinical judgment In nursing. *Journal of Nursing Education, 45*(6), 204-211.

Tanner, C.A., Benner, P., Chesla, C, & Gordon, D.R. (1993). The phenomenology of knowing the patient. *Image: Journal of Nursing Scholarship, 25*(4), 273-280.

Texas Board of Nursing, (2011). *Position statement 15.28: The registered nurse scope of practice.* Retrieved from http://www.bon.texas.gov/practice_bon_position_statements_content.asp#15.28

Texas Board of Nursing, (2011). *Differentiated Essential Competencies of Graduates of Texas Nursing Programs.* Retrieved from https://www.bon.texas.gov/pdfs/publication_pdfs/delc-2010.pdf

Todd, M., Manz, J.A., Hawkins, K.S., Parsons, M.E., & Hercinger, M. (2008). The development of a quantitative evaluation tool for simulations in nursing education. *International Journal of Nursing Education Scholarship, 5*(1), 1-17.

Traynor, M. & Galanouli, D. (2015). Have OSCEs come of age in nursing education? *British Journal of Nursing, 2015, 24*(7), 388-391.

Viera, A.J. & Garrett, J.M. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine, 37*(5), 360-363.

Vinales, J.J. (2015). Exploring failure to fail in pre-registration nursing. *British Journal of Nursing, 24*(5), 284-288.

Voepel-Lewis, T., Pechlavanidis, E., Burke, C., & Talsma, A.N. (2013). Nursing surveillance moderates the relationship between staffing levels and pediatric postoperative serious adverse events: A nested case-control study. *International Journal of Nursing Studies, 50,* 905-913.

Walsh, M., Bailey, P.H., Mossey, S., & Koren,, I. (2010). The novice objective structured clinical evaluation tool: psychometric testing. *Journal of Advanced Nursing, 66*(12), 2807-2818.

Walshe, N., O'Brien, S., Hartigan, I., Murphy, S. & Graham, R. (2014). Simulation performance evaluation of the DARE^2 –Patient Safety Rubric. *Clinical Simulation in Nursing, 10*, 446-454.

Waltz, C.F., Strickland, O.L. & Lentz, E.R. (2005). *Measurement in nursing and health research, (3d. ed.).* New York: Springer Publishing Company.

Watson, R., Stimpson, A., Topping, A., and Porock, D. (2002). Clinical competence assessment in nursing: a systematic review of the literature. *Journal of Advanced Nursing, 39*(5), 421-431.

Whyte, J. & Cormier, E. (2014). A Deliberate Practice-Based Training Protocol for Student Nurses: Care of the Critically Ill Patient: A Randomized Controlled Trial of a Deliberate Practice-Based Training Protocol. *Clinical Simulation in Nursing, 10,* 617-625.

Wolf, L., Dion, K., Lamoureaux, E., Kenny, C., Cumin, M., Hogan, M.A., Roche, J., & Cunningham, H. (2011). Using simulated clinical scenarios to evaluate student performance. *Nurse Educator, 36*(3), 128-134.

Biographical Information

Tamara Andrews has been a nurse for 29 years. She has worked in the areas of medical-surgical nursing, neurosurgical nursing, neurosurgical intensive care, women's and children's services, psychiatric nursing, and nursing education. Dr. Andrews worked as a hospital-based nurse educator for 11 years, and a university professor for 12 years. Her research interests include nursing education, educational strategies and curriculum, and faith-based nursing. She holds a Master's Degree in Nursing from the University of Texas at Arlington. Dr. Andrews served on the Advisory Board for the Lippincott, Williams Wilkins/Laerdal project for the development of a virtual simulation product for Health Assessment.

She has taken a position as Associate Professor of Nursing at the University of Mary-Hardin Baylor.