A Systematic Review of the Validity and Reliability of Observational Measurement Tools Evaluating Undergraduate Nursing Students' Individual Competency Outcomes Following High Fidelity Simulation-Based Clinical Experiences

by

Barbara A. Bleasdale

Submitted in partial fulfilment of the requirements
for the degree of Master of Nursing

at

Dalhousie University
Halifax, Nova Scotia
June 2015

Dedication Page

This work is dedicated to my parents, the late Robert P. Boutilier who loved to read and learn and encouraged his children to seek higher education and who knew they would be successful even when they had doubts; and to my mother, Barbara A. Boutilier who encouraged us to do our best and stood beside us all the way and continues to unconditionally love and support us to this day.

The thesis is also dedicated to two of my dear nursing colleagues and mentors who passed too soon: Colleen Kiberd and Cynthia Barkhouse-MacKeen. Both encouraged me to pursue this degree and topic and gave me invaluable advice and support. Colleen and Cynthia were excellent teachers and clinicians who were passionate about nursing, students and nursing education and raised the bar for those of us who follow them.

It is my hope that this work brings honour to each one of you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Competence outcomes of simulation-based clinical experiences (SBCEs) are being reported in the literature but it is not clear if current SBCE tools use observational measures of competence outcomes nor is the instrument's validity and reliability clearly established for use with undergraduate nursing students following high fidelity SBCEs. This systematic review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines and searched the nursing literature (open start date - November 2014) focusing on SBCE outcomes of nursing competence, clinical judgment, clinical reasoning, and critical thinking. Nineteen studies were included and results indicated three tools that met standard reliability and validity criteria, and another three tools that met specific quality assessment criteria. These results raise the questions: how is nursing competence defined, what level of performance indicates competence for nursing students and graduate nurses and, is more than one tool required to accurately and comprehensively measure student competence performance outcomes.

LIST OF ABBREVIATIONS USED

AD	Associate Degree
BN	Baccalaureate Degree
PN	Practical Nurse
HPS	Human Patient Simulator
JR	Junior
SBCE	Simulation-Based Clinical Experience

ACKNOWLEDGEMENTS

I would like to acknowledge my supervisor, Dr. Audrey Steenbeek whose research expertise inspired me to strive for excellence in both methodology and writing and who generously kept me on as her student despite a major change in my thesis topic and methodology. I extend to her my thanks for her time, expert review and support throughout this journey.

As well, I extend my deep appreciation to my committee members Dr. Shelley Cobbett, Prof. Colette Foisy Doll, and Prof. Kim Hebert. Your expertise in simulation and education brought depth and value to my thesis and your thorough and timely review of my work was always appreciated.

I also acknowledge the support, work and in depth review provided by my second reviewers: Dr. Lisa Adams and Chola Shamputa. Thank you Chola for your assistance in creating my study selection forms, and thank you Dr. Adams for stepping up and offering me your help at a critical point in my review. Your help allowed me to continue with my study and I learned so much from our discussions and your review of the studies.

Last, but by no means least, I acknowledge the ongoing and steadfast support of my family throughout these past six years as they dealt with my ups and downs and always told me to keep going; "you can do it!" To my son, Adam- thank you for your understanding and keeping the home fires burning while I was away for many months; to my daughter, Laura and son-in-law, John – thank you for your pep talks and bringing me the joy of your wedding last year, and to my partner of many years, Roger thank you for believing in me and encouraging me. Thank you to my grandchildren Marissa, Kyla, Eva, Mattias and Adam for asking me how my paper was going and understanding why Nanny couldn't be with you as often as we would have liked.

I also thank my Mum for her help with filing, sorting and organizing so many articles, and for supporting and listening to me night after night and to my brother Bob for his support and encouragement. Finally, I owe a debt of gratitude and offer my thanks to my sister, Judy and brother-in-law, Roy for making me welcome in their home at critical writing times and keeping me calm and focused! Thank you Judy for giving me the push I needed when I was tired, and for your dedicated and excellent assistance in reading and editing my chapters so carefully. Thank you Roy for providing IT support on-call to me- it was always appreciated.

Thanks to each of you for your love, patience, for standing beside me and reminding me of what is possible and for your understanding each time I was away from important family events to finish the next draft!

CHAPTER 1 INTRODUCTION

Competent practitioners are fundamental to ensuring the safety of patients who require and expect competent care from health care providers including nurses and nursing students (Wolf et al., 2011). However, due to numerous factors, it is becoming increasingly difficult to ensure that students graduate with an acceptable level of competence. These include: nursing shortages, increasing levels of patient acuity, the increasingly complex and unpredictable nature of current health care environments, shorter hospital stays, increased enrollment of nursing students, and decreased acceptance of students "practicing" on patients (Carlson, 2011; Galloway, 2009; Rhodes & Curran, 2005; Ziv, Wolpe, Small, & Glick, 2003). These same factors have made it steadily more difficult to secure clinical practice placements in healthcare agencies (Nehring & Lashley, 2004). This is of particular importance as the lack of clinical site practice has direct impact on student competence development due to fewer opportunities to experience and act in complex care environments and in high risk situations.

Clinical reasoning is a key component of competence (based in experiential learning) which occurs during clinical practice, and the literature shows that nurses with effective clinical reasoning skills have a positive impact on patient outcomes while nurses with poor skills in this area fail to detect impending patient deterioration, thus compromising patient safety (Lapkin, Levett-Jones, Bellchambers, & Fernandez, 2010). With few opportunities to develop clinical reasoning, new graduates may not be as well prepared to meet the challenges that await them upon entry into the nursing profession (Norman, 2012). As such, it is essential that nursing education programs include alternate ways of providing clinical experience and an associated method of evaluation of their outcomes.

Nursing is a practice-based profession, and traditionally clinical nursing education was based on an apprenticeship model of learning whereby students were supervised in their skill sets during patient care delivery in various clinical settings (Oldenburg, Maney, & Plonczynski, 2013). In these situations, nursing competence is determined by measuring a student's knowledge, skills, and attitude in providing safe patient care (Meakim et al., 2013). Essentially, it is in the clinical area that student nurse competence is evaluated by the instructor (or preceptor) through observation of the students' performance and decision making skills (referred to as clinical reasoning) (Tanner, 2006) while providing patient care (Cowan, Norman, & Coopamah, 2007). Currently, this apprenticeship model of active learning continues to be the preferred method for nursing competency achievement (Sportsman et al., 2009).

Clinical Site Evaluation Concerns

There are problems associated with evaluating student competence in the clinical area (Isaacson & Stacy, 2009). Students' concerns with clinical performance evaluation include anxiety, uncertainty regarding completing self-evaluations, general grade expectations, and various grading standards by members of the clinical team (Isaacson & Stacy). Instructors' concerns with clinical performance evaluation include the subjectivity of the evaluation experience, confusing jargon in evaluation tools (Fahy et al., 2011), and challenges related to the overall inability to control the clinical environment in order to provide specific experiences for students (Isaacson & Stacy). Instructors often struggle with displaying fairness and consistency (Fahy et al.) in student evaluations in various situations that may include: the student who does not have the opportunity to show their capabilities based on their clinical patient assignment; determining how much of a student's performance evaluation should be based on whether or not the unit was able to provide an appropriate learning environment; and, whether or not the impact of staff behaviour has affected the student's competence in learning experiences (Isaacson & Stacy).

Other concerns regarding on-site evaluation include the lack of time available for clinical site instructors to think through clinical problems with students to encourage student clinical reasoning skills (Lapkin et al., 2010) and the difficulty in assessing a student's reactions to changes in a patient's condition when the student has been asked to step aside from the emergent situation while the experienced nurse takes over the situation (Jensen, 2013). A compounding factor is that novice clinical instructors are often used by nursing programs for formal evaluation of student competence in the clinical area. Novice clinical instructors' unfamiliarity with academic evaluation methods and uncertainty surrounding their role of being both mentor and evaluator at different times in one experience can lead to problems in the completion and validity of student evaluations (Isaacson & Stacy, 2009).

The use of simulated clinical experiences may provide a solution to the lack of consistent complex care clinical practice opportunities as well as address the issues surrounding clinical site evaluation (Isaacson & Stacy 2009). Simulation-based clinical experiences (SBCE) can be a valuable tool in providing nursing students with consistent clinical learning environments, complex patient assignments, clinical testing environments, and evaluators who are familiar with evaluation tools and methods

(Brewer, 2011; Buykx et al., 2011). Norman (2012) states that employing simulation methods which are conducive to experiencing real life healthcare situations would ultimately help students achieve improved and more competent outcomes in health care delivery, thus better preparing new graduates to enter complex environments. In fact, one of the major benefits of simulation training is reported to be improved client safety and better prepared nurses (Garrett, MacPhee, & Jackson, 2010).

Simulation-Based Clinical Experiences (SBCE)

A simulation-based clinical experience is an activity that mimics the reality of a clinical environment and is designed to demonstrate procedures, decision-making, and critical thinking through techniques such as role playing and the use of devises such as interactive mannequins (or videos) (Jeffries, 2005). The student participates in a patient care situation via the simulated clinical experience which takes place in a simulation lab. The features of the patient scenario and environment must be authentic and include as many realistic environmental factors as possible (Jeffries). This is based on Gaba's (2004) definition of simulation as "a technique, not a technology, to replace or amplify real experiences with guided experience, often immersive in nature, that evoke or replicate substantial aspects of the real work in a fully interactive fashion" (p. 2).

Researchers have found that high fidelity patient simulation-based learning outcomes are equal to or better than those of other teaching methods including practice in an actual patient situation (Sportsman et al., 2009). Clinical simulation experiences provide enhanced skill performance, increased clinical knowledge, and more refined critical thinking abilities as possible learning outcomes for the student (Bland, Topping, & Wood, 2011). This methodology is also supported by the Canadian Patient Safety

Institute (CPSI) which recognizes that the end goals of simulation in healthcare are to improve performance of health professionals, reduce human error and increase patient safety (CPSI, 2008). Indeed, simulation has become an established pedagogy for teaching clinical skills and now forms a significant part of the curriculum in undergraduate, graduate, and continuing professional education for most health professionals in Canada (Bland et al., 2011; Canadian Patient Safety Institute, 2008; Garrett et al., 2010).

Background

History of simulation.

Simulation has been used by various professions in the education of students, including aviation, military, nursing, and medicine. In the early days, simulation techniques in healthcare included practice on cadavers, mannequins, and small anatomical models. In 1910 the nursing manikin called *Mrs. Chase* was created, believing that a manikin would offer nursing students an opportunity to put theories into the practice of clinical nursing skills. The aviation industry began using simulation with the invention of the blue box flight trainer in 1929 which has been used as a training tool in the civil air industry since 1955. It was this industry's technology which gave rise to the first full body patient simulators for anesthesiology education in the 1970s (Nehring & Lashley, 2004). Aviation and military education have continued to refine and use high fidelity simulation for all student training as they found savings in cost and lives with this methodology (Rosen, 2008).

Various task trainers (e.g. intravenous arms, catheterization models) became available for specific skill practice in health professional education, but it was not until the 1990s when computerized patient manikins were available in nursing education.

Basic patient simulators are computerized manikins with factory installed human qualities such as heart, lung, and voice sounds, but are not fully interactive so are not used for high fidelity experiences. The manikins used in high fidelity clinical simulations are currently referred to by various names: computerized patient simulators (Horan, 2009), human patient simulators, patient simulators, and high fidelity patient simulator (Jeffries, 2007). High fidelity patient simulators (HFPS) can be programmed to speak, breathe, have palpable pulse, audible breath and heart sounds, and react appropriately to medications and defibrillation in order to deliver countless clinical practice scenarios.

Benefits of Simulation-Based Clinical Experiences

SBCE using high fidelity patient simulators has several benefits for nursing students that include: the focus is placed on student learning rather than on the patient, as is the case for clinically- based practice experiences (Decker, Sportsman, Puetz, & Billings, 2008; Ziv et al., 2003); it offers students a variety of clinical problems and practice with associated clinical reasoning skills (Kuiper, Heinrich, Matthias, Graham, & Bell-Kotwall, 2008); it allows learners to practice in a relaxed environment where there is no pressure to perform quickly and accurately without mistakes (Brewer, 2011); and, it allows consequences of mistakes to occur so students may learn from these mistakes (Jeffries, 2007). SBCE appeals to technology savvy students while at the same time providing for nonlinear thinking, which is more familiar to them, and helps them to retain information (as it requires coordination of cognitive, affective, and psychomotor skills) (Spunt, 2004 as cited in Starkweather & Kardong- Edgren, 2008). The utilization of non-linear thinking and the greater ability to retain information is true for all students experiencing simulation because teaching and learning principles of simulation are based in experiential learning.

Experiential learning is based on student participation in the learning experience, and it is this participation that helps the student apply psychomotor and cognitive skills as they would in the actual practice site (Alinier, Hunt, & Gordon, 2004; Jeffries, 2007). A further benefit of simulation is that learning is contextualized in an environment similar to that in which the graduate will practice and these contextualized learning experiences seem to lie at the base of professional learning (Kneebone, Scott, Darzi, & Horrocks, 2004; Schuwirth & van der Vleuten, 2003). Alinier, Hunt, Gordon, and Harwood, (2006) suggest that in the future, nursing graduates may be expected to be competent in handling clinical emergencies after having practiced primarily with human patient simulators.

Competence measurement

In a discussion regarding the measurement of competence outcomes it is necessary to define the term competence as understood by the nursing profession. Although there is no consensus on the definition of competence in the nursing literature (Cowan et al., 2007; Fahy et al., 2011; Smith, 2012), a definition that is in current use by the simulation community describes competence as a combination of discrete and measurable knowledge, skills, and attitudes that are essential for patient safety and quality of care (Meakim et al., 2013). The skills in this definition of competence include psychomotor and affective skills as well as the cognitive and meta-cognitive skills of critical thinking, clinical judgment, and clinical reasoning. Critical thinking, clinical judgment and clinical reasoning are frequently used interchangeably in nursing research literature. Clinical reasoning is identified as a key component of competence because capable professional practice requires complex thinking processes as well as knowledge and psychomotor and affective skills (Banning, 2008).

Nursing scholars have carried out research to better understand the degree to which the simulated clinical experience lends itself to the attainment of clinical competencies (Childs & Sepples, 2006; Frontiero & Glynn, 2012; Horan, 2009; Jeffries, 2005; Sportsman et al., 2009) and critical and reflective thinking skills as they are an essential element in providing safe patient care (Alinier et al., 2004, 2006; Decker et al., 2008; Gaba, 2004 **;** Garret et al., 2010; Moule, Wilford, Sales, & Lockyer, 2008; Seropian, 2003; Starkweather & Kardong- Edgren, 2008). Jensen (2013) notes that the simulated environment is one in which students' clinical reasoning skills may be evaluated as well as fostered. Lasater (2007b) carried out research on SBCE outcomes and also determined that simulation is useful in the development of clinical judgment which is critical in competent nursing care. The outcomes of clinical judgment are: the ability to recognize changes and salient aspects in a clinical situation; to interpret their meaning; to respond appropriately; and, to reflect on the effectiveness of the intervention (Meakim et al., 2013).

Outcomes of Simulation-Based Clinical Experiences

With the goal of the simulated clinical experience being to improve or change knowledge, skills, and attitudes (KSA) of the learner, it is essential to define the outcomes expected from an SBCE event. Outcomes in SBCE are defined as the measurable results of the participant's progress toward meeting a set of objectives. Expected outcomes are the change in KSA as a result of the simulated experience (Meakim et al., 2013). There is continuing research on the outcomes of simulation-based clinical experiences, their measurement, and a move forward in the use of simulation as an effective evaluation strategy (Frontiero & Glynn, 2012; Radhakrishnan, Roche & Cunningham, 2007; Sportsman et al., 2009; Todd, Manz, Hawkins, Parsons & Hercinger, 2008).

Several researchers have measured SBCE outcomes using student self-assessments to evaluate competence and self-confidence outcomes (Bambini, Washburn & Perkins, 2009; Barnsley et al., 2004; Baxter & Norman, 2011; Lauder et al., 2008; Smith & Roehrs, 2009). Kardong-Edgren et al. (2010) reported that self-assessment measures have not been correlated with tester perceptions; in fact, the least skilled practitioners were shown to be the most confident. Cant and Cooper (2010) encourage research on outcomes measurement of actual student performance noting that proxy measures such as perceived competence and knowledge may not be a valid indicator of performance skills.

A quantitative pilot study (Radhakrishnan et al., 2007) indicated that clinical performance parameters are influenced by HFPS practice and that student performance in simulated clinical experiences is measurable. In their 2008 study, Todd et al. measured student performance in a quantitative study and created a quantitative evaluation tool to assess student performance in clinical simulation experiences and identified 22 behaviours associated with critical thinking, communication, assessment, and technical skills. Clearly it is necessary to have valid and reliable observational measurement tools that can determine clinical competence outcomes from SBCE.

Problem Statement

Competence outcomes of SBCE are being reported in the literature but it is not clear if current SBCE evaluation tools use observational measures of competence outcomes, nor their validity and reliability for use with undergraduate nursing students following high fidelity simulation-based clinical experiences. This is due in part to the fact that each researcher focuses on specific and different competence outcomes based on the selected definition of competence and, more often than not, each employs a different measurement instrument and population. As such there is a lack of continuity of testing with one tool that, if done, would ultimately provide evidence for nurse educators to determine the best measure of student competence outcomes following simulated clinical events.

Often, competence is the identified outcome indicator for measurement while others report different indicators for competence measurement. The competence and competence assessment literature was reviewed for the years 2000-2007 (on behalf of the National Cancer Nursing Education Project Australia (EdCan) 2008) in an effort to identify the best available evidence related to competence assessment tools and processes; it is largely accepted that more than one indicator should be used in the assessment of competence. Clinical judgment, clinical reasoning, critical thinking, and competence are all indicators used frequently in research measuring competence outcomes, and so this study will consider all four terms as competence outcomes and endeavor to capture all observational measurement tools in use.

Adamson and Kardong-Edgren (2012) noted that a major obstacle to accurate evaluation of learning outcomes is a lack of evaluation instruments that allow nurse educators to make valid and reliable evaluations of student performance in high fidelity simulation activities. Subjectivity of tester ratings has been noted as a problem with both clinical site and simulation on-site competence evaluations (EdCan, 2008). Currently there is a call for objective measures of competence (Oldenburg et al., 2013) and more specifically, for a standardized method to quantitatively measure students' performance outcomes (Frontiero & Glynn, 2012). A systematic review of the literature seeking evidence of valid and reliable observational measures of nursing student competence outcomes following high fidelity simulation clinical experiences will hopefully identify objective testing methods with quantitative measurements. Definitions of the terms valid and reliable as they are applied to this study appear in Appendix A.

Gaps in the Literature

Several gaps are identified in the literature regarding the measurement of competence outcomes within a simulated clinical experience. First, there is a lack of consensus on the definition and meaning of the term competence as it relates to the profession of nursing and exactly what constitutes critical thinking and competence(y) (Cowan et al., 2007; Fahy et al., 2011; Smith, 2012). Smith notes that without a clear definition of competence it is difficult to identify how nurses develop competence, and subsequently establish methods of evaluation for this purpose.

Second, there is a lack of clarity on what competence outcomes are and how they should be measured. A systematic review of the literature for the years 2000-2010 focused on simulation outcomes in nursing education (Norman, 2012). Norman found outcome measurements included knowledge and skills, safety, communication, clinical judgment, satisfaction, confidence, and clinical evaluation. She further categorized these outcome measurements into three themes: external outcomes, internal outcomes and evaluation of outcomes. Internal outcomes were identified as those that were dependent on learners' insights such as clinical judgment, satisfaction and self-confidence. External outcomes were identified as factors that are learned: knowledge, skills, safety, and communication outcomes. Evaluation of internal and external outcomes can be used to determine clinical performance levels of nursing students. Finally, within the research on existing and new measurement tools it is not always clear which competence outcomes they measure, whether or not they are observational measures of competence and whether or not the study design met standards of rigour.

Kardong-Edgren, Adamson and Fitzgerald (2010) cautioned researchers of new studies using existing evaluation tools to ensure that the tool they select is a valid and reliable measure for their population, participants and raters. The authors state that researchers will assist in the growth of simulation pedagogy by: aspiring to higher levels of evaluation, reporting psychometric measures and by taking steps to assure validation with new populations. Examples of these tools include grading rubrics (Gantt, 2010; Lasater, 2007; Morgan & Cleave-Hogg, 2002), simulation evaluation tools developed by faculty (Radhakrishnan et al., 2007; Todd et al., 2008), objective structured clinical examinations (OSCE), (Alinier, Hunt, Borden & Harwood, 2006; Baxter & Norman, 2011; Lauder et al., 2008), and pretest/post-test of knowledge acquisition (Schlairet & Pollock, 2010).

A review of published simulation evaluation instruments in 2010 (Kardong-Edgren et al.) categorized tools by: 1. author(s) name, 2. instrument, 3. validity, 4. reliability results, 5. learning domain (including comments), and stressed that no new instruments should be added to the current research cadre until the current ones were tested. In 2013, this review was updated (Adamson, Kardong-Edgren, & Willhaus, 2013) to include a current and expanded list of instruments which were reviewed under the same categories. The authors also describe two frameworks for categorizing simulation evaluation strategies.

Study Aim

The purpose of this study is to carry out a systematic review of the literature from open start date-2014 that is focused on simulation-based clinical experience outcomes of nursing competence, clinical judgment, clinical reasoning, and critical thinking, and to determine which tools provide observational outcome measures that are both valid and reliable. This study is based on the assumption that these outcomes can be measured by observation.

To address this aim the research question will be:

What observational measurement tools are reliable and valid in measuring students' individual competency outcomes for simulation-based clinical experiences using high fidelity patient simulators for undergraduate nursing student clinical experiences?

Method

A systematic review of the literature will be conducted to identify, select, critically appraise, and synthesize high quality primary research studies that are relevant to this question while adhering to explicit guidelines for the conduct of a quantitative systematic review of the literature. The evidence selected will meet pre-set criteria for inclusion and exclusion (contained in a selected appraisal tool) to ensure that the design of the selected studies is explicit and rigorous. Once the studies have been appraised the researcher will analyze the selected study results for credibility to evaluate the strength of the evidence. The study characteristics, quality and results will be presented in tables. Finally, the collected material will be organized by study design and combined by narrative methods to present a critical analysis of the findings. Descriptive statistical findings will be included in summary charts.

The search will be comprehensive, including electronic databases, searching references in selected studies and searching for studies with more limited distribution (grey literature) as in simulation conference papers and dissertations between the years 2000- early 2014.

Significance to the Nursing Profession

This study will provide the nursing community with an overview of the nursing simulation evaluation literature by offering an in-depth review and critical analysis of selected studies, competence outcomes measured by observational assessment and identifying which of the measurement tools, if any, are valid and reliable. A systematic review of the literature with the inherent standards and structured scientific method of enquiry will provide an appraisal and synthesis of all studies relevant to the research question despite their results. The value of a systematic review for policy, practice decisions, education and further research lies in the methodology which includes: the use of transparent and explicit methods determined before the search begins; an adherence to stage by stage search, analysis and synthesis of the literature; and, a review that is replicable and able to be updated due to the transparent reporting of methods (Houde, 2009).

These strategies seek to avoid bias in the selection of primary research and their results, thus rendering an encompassing and objective view of the literature in relation to the question. Because a systematic review reflects all relevant, scientifically sound research it is able to present a balanced view of the evidence (Petticrew & Roberts, 2006). Therefore, a systematic review is considered an excellent method to present evidence for nursing practice, education and research (Holopainen, Hakulinen-Viitanen, & Tossavainen, 2008). The results of this study seek to provide evidence towards the need for evidence-based evaluation and practice in nursing education and the requirement for psychometrically sound evaluation instruments to support evidence-based teaching practices.

The significance of developing evaluation measures in evidence-based practice is two-fold. The first is that evidence-based, psychometrically sound evaluation instruments will enable researchers to conduct rigorous research about current and future educational practices and second the data produced from these studies will allow nursing education scholars to make decisions to improve (teaching and learning in simulation) in the future (Adamson & Kardong-Edgren, 2012).

Simulation and Related Terminology

In order to communicate clearly and provide guidance regarding simulation and related terminology used in this paper, definitions are presented in Appendix A. The terms presented will be used in a consistent manner in all aspects of the study to promote understanding between researcher and reader. The simulation definitions are taken from the Standards of Best Practice: Simulation contained in *Clinical Simulation in Nursing* (2013) (Standard 1: terminology) and were developed by leaders in the nursing simulation community. These definitions were chosen as they have been adopted by simulation centers in the USA and internationally, and are frequently cited in the simulation literature and used for developing simulation research proposals (Meakim et al., 2013). Terms that were not defined in the simulation terminology standard (Meakim et al., 2013) or were not defined appropriately for this study are cited with their sources in Appendix A.

Summary

The proposed research will examine the criteria for establishing validity and reliability in tools designed to measure observable competency outcomes in high fidelity SBCE. Additionally, background information of simulation and its' ability to present options for competency learning and evaluation of students in increasingly unavailable clinical sites will also be presented. Chapter two will focus on a review of the literature pertinent to this study and will include: a presentation of current measurement tools and associated pros and cons; a discussion of the meaning of the word competence as used in nursing education literature along with its associated concepts of clinical judgment, critical thinking and clinical reasoning. This discussion will include an overview of relevant literature examining how SBCE lends itself to measurement of the outcomes of these concepts.

Chapter 3 will present the methodologies to be incorporated in the systematic review including the review protocol, the search strategy, the selected appraisal tool with the inclusion and exclusion criteria for selecting studies, the data analysis methods for categorizing the data and for analyzing the primary studies for the strength of their evidence, and the method of presenting the results by narrative and tables. Chapter 4 will present the results. Finally, Chapter 5 will provide a discussion of the findings that will include a critique of the primary study findings as it relates to the research question by identifying tools that are observational measures of competence, their validity and reliability status, the specific outcomes that were measured, and a tabular presentation of included studies as well as descriptive statistics.

CHAPTER 2 LITERATURE REVIEW

The objective of this literature review is the validity and reliability of currently used competence measurement instruments in SBCE. In addressing competence measurement it is necessary to have a clear understanding of competence, thus this chapter first examines the discrepancies that exist around competence as it relates to the profession of nursing. Additionally, critical thinking (CT), clinical judgment (CJ), and clinical reasoning (CR) terms, often associated with competence, are also explored. Finally, the definition of competence that will be used in this review and accompanying rationale will conclude this discussion.

In measuring competence it is also necessary to identify outcomes and measurement strategies. Therefore, research studies that address competence outcomes and how they are measured; as well as the outcomes of CT, CR, and CJ and related measurement tools. Rationale for their inclusion is provided. Lastly, a discussion of standards of rigour associated with measurement tools and ability to measure competence through observation is presented. Chapter two concludes with a brief discussion of the systematic review method and rationale for selecting this method for the proposed study.

Introduction

A literature review was conducted to examine competence measurement in simulated clinical exercises with students. The search was conducted via several electronic databases: CINAHL, Medline/ PubMed, EMBASE, and citations in bibliographies. Inclusion criteria were: English only articles and books in health sciences, aviation, and military for the years 2000 – early October 2013 using the broad search terms nurs*, competence, simulation, measurement, validity, student, evaluation, and instruments. The search was further the refined by using the terms “high fidelity simulation” AND measurement AND student competence.

Four major themes became apparent: (a) simulation as an educational evaluation strategy; (b) improved patient safety and quality of care resulting from learners taught and evaluated through simulation techniques; (c) ethical considerations surrounding the use of simulation-based education for competency development and evaluation; and (d) lack of clarity regarding instruments’ validity and reliability in measuring competence related outcomes.

This study focuses on the fourth theme, that of competence measurement and within this theme three issues will be further discussed as follows: (a) lack of a consistent definition of competence and related outcomes; (b) varying evidence on the ability of SBCE to measure competence and, (c) a lack of clarity on which observational measurement tools (i.e. in current use with SBCE and high fidelity patient simulators) are a valid and reliable measure of undergraduate students’ individual competency outcomes.

Competence in Nursing

The challenge of establishing competencies for measurement begins with the lack of consensus regarding the definition of competence as it is used in nursing. The literature reflects this lack of consensus as, currently, some studies make reference to the term competence but do not provide an operational definition while other studies use other terms to reflect competence (e.g. student performance; skills performance) (Alinier et al., 2006; Baxter & Norman, 2011; Lauder et al., 2008). There are several suggested reasons for this lack of consensus on the definition of competence, beginning with the confusion between the terms competence and competency. There are two divergent views about these definitions: (a) that competence is an aspect of a job that a person performs while competency is the behaviour underpinning the performance and, conversely, (b) that competence is the knowledge, capacity and potential to perform skills while competency is the actual performance in accordance with policies in a situation (Cowan et al., 2007). While there is no consensus on the definitions of competence and competency, Benner's (1982) definition of nursing competence as "the ability to perform a task with desirable outcomes under the varied circumstances of the real world" is frequently quoted in nursing literature and incorporates capability and performance in one term (Cowan et al.).

Watson (2002) suggests that competence is simply the lack of "incompetence" and that competence may not be the most optimal benchmark nursing should measure itself against. She believes that this focus would make the profession task-oriented and thereby hinder nurses' educational and professional development, as the main focus would be developing competence in tasks rather than developing higher level attributes. Cowan et al. (2007) argue that another factor that creates a challenge for defining nursing competence is that nursing requires a complex blend of knowledge, skills, performance, and attitudes which are difficult to capture in one concept.

There are other factors that also impact the selection of a definition for nursing competence. Cowan et al. (2007) note that the move away from the apprenticeship model of nursing education to one that involves institutions of higher learning in the year 2000, led to research-based nursing practice and a move away from meeting simplistic levels of practice such as standards. This move created some tension within health care practice and service sectors, which typically want nursing graduates that meet basic standards and are ready to practice with minimal orientation time. Cowan et al. argue that when competence was restated in broad terms like: assess, plan, implement, and evaluate, the apprenticeship practical procedural model of learning was excluded, which in turn resulted in unmet service needs. This occurred during a time frame when academic and personal development became the main focus of education rather than only practical procedural learning. Today, the role of the nurse is more diverse and is moving away from the historically defined role of caring for the ill, thus creating a challenge for the service sector to define these new roles of competence in practical competency statements (Smith, 2012).

Nursing regulatory bodies want a clear definition of competence in order to define the requirements for a new graduate nurse. Competence stated as competencies is preferred so regulatory bodies may communicate the requirements clearly with nursing educational institutions (Yanhua & Watson, 2011). It is obvious that there are many reasons why defining competence is difficult, and at times problematic, yet there is a consensus that competence must be defined and measured (Cowan et al., 2007).

Concepts of Competence

Nursing scholars discuss three concepts of competence: behaviourist, generic and holistic (Cowan et al., 2007; Garside & Nhemachena, 2013). The behaviorist or performance concept of competence (as defined by Gonczi, 1994 in Cowan et al.) is task-based, describing nurse's discrete behaviours associated with task performance. This concept makes the task synonymous with competency, while ignoring any connections between various tasks and the nurse's attributes (Cowan et al.). The behaviourist concept is based on discrete behaviours associated with the individual's level of performance (i.e. tasks and skills), as such it is criticized as being reductionist as it excludes the role of professional judgment, which is required in real-world complex care environments (Cowan et al.). Alternatively, Fahy et al. (2011) note that the behaviourist approach is broad, combining a psychological component and the ability to perform tasks that include affective, cognitive and psychomotor skills.

The generic concept of competence is described as person-oriented including the underlying characteristics and qualities of the individual as indicators of effective performance (Cowan et al., 2007; Fahy et al., 2011). These indicators include the transferable attributes of knowledge, problem-solving and critical thinking capacity. The generic concept of competence is often criticized for ignoring that nursing practice is context-dependent whereby the nurse’s scope of competence is related to the clinical context in which care is given, and as such, the level of competence may change when moving from one specialty area to another (Garside & Nhemachena, 2013).

The holistic integrated concept of competence is inclusive of the general attributes of the nurse and the practice context drawing on the nurse’s knowledge, skills, attitudes, values, and professional judgment for effective performance (Fahy et al., 2011). The inclusion of the nurses’ attributes and the context of practice should then include ethics, values and reflective practice in the components of competence. The holistic concept is inclusive of both the behaviourist and generic concepts and enables assessment of the practitioner’s capacity to integrate them in their practice (Gonczi, 1994, in Cowan et al., 2007).

One study that identifies a definition of competence is Decker et al. (2008): “the acquisition of relevant knowledge, the development of psychomotor skills, and the ability to apply the knowledge and skills appropriately in a given situation” (p. 75). This definition incorporates the behaviourist concept of task development by including psychomotor skill development. Additionally, the definition incorporates the generic concept stressing the importance of the nurse's abilities and attributes by including the nurse's ability to apply the knowledge and skills appropriately.

Garside and Nhemachena (2013) note that as nursing includes a diversity of dimensions that cannot be readily reduced to a mechanistic list of competencies, they support the holistic approach as it identifies broad groups of general attributes considered essential for effective performance and provides a basis for transferable skills in delivering care and tools to measure it. Cowan et al. (2007) state that by accepting the more encompassing holistic approach it could facilitate greater acceptance of the competence concept and provide researchers with the definition needed to establish competence standards.

Selected Definition of Competence

As noted earlier, many regulatory bodies both national and international have tried to reach consensus on the definition of competence (Yanhua & Watson, 2011). In 2011-2012 Canada's ten nursing regulatory bodies including the College of Registered Nurses of Nova Scotia (CRNNS) cooperatively define competence as "the ability of the registered nurse to integrate and apply the knowledge, skills, judgments and personal attributes required to practice safely and ethically in a designated role and setting" (Black, Allen, & Redfern, 2008, p. 173). To date CRNNS continues to use this definition of competence as it: "... includes both entry-level and continuing competencies" (p. 14) (CRNNS, 2012).

In 2013, the International Nursing Association for Clinical Simulation and Learning (INACSL) define competence in their Standard 1 Terminology as "a combination of discrete and measurable knowledge, skills, and attitudes that are essential for patient safety and quality patient care" (Meakim et al., 2013 p. S5). INACSL created the Standards of Best Practice for Simulation which have been cited in publications and used in research and funding proposals as well as in designing and implementing simulation experiences. This study uses the INACSL definition of competence as it is reflective of the holistic concept of competence as recommended by authors in this literature review as well as the definitions presented by national and local regulatory bodies.

Critical Thinking, Clinical Judgment, Clinical Reasoning

The terms critical thinking (CT), clinical judgment (CJ), and clinical reasoning (CR) are frequently included in discussions around competence and the literature shows that there is no consensus on their definition in nursing. In the following discussion each term is presented separately with a description of its' current use and definitions and its relationship to the concept of competence. The selected definition for use with this systematic review is also provided.

Critical thinking.

While there is no consensus on the definition of CT, most CT definitions are similar (Ravert, 2008). Smith (2012) in her concept analysis of nurse competence, discusses CT as a higher-level cognitive function using knowledge, prior experience, judgment, reasoning, and analysis to provide effective individualized nursing care. Meakim et al. (2013) defines CT as a methodical process whereby the nurse validates health care data while noting any personal or professional assumptions that may influence her or his thoughts and actions. This is followed by reflecting on the entire process while considering the effectiveness of what has been determined as the necessary action(s) to take. Similarly, Brunt (2005a) notes that CT skills enable the nurse to consider various possibilities in a clinical situation; consider alternative problems and interventions; weigh the consequences of each option and select the most appropriate action. Victor-Chmil and Larew (2013) describe CT as a cognitive process used to analyze knowledge based on evidence and science and emphasized that CT is not a discipline-specific skill.

There does appear to be consensus in the nursing literature that CT is a process of purposeful thinking and reflective reasoning in the context of nursing practice which is associated with a spirit of inquiry, logical reasoning, discrimination, and the application of standards (Brunt, 2005b). As nurses move along the competence continuum from novice to expert, they improve their ability to think critically (Brunt, 2005a). It is important to note that there is agreement in the literature that critical thinking is considered a core competency for the professional nurse (Ravert, 2008; Smith, 2012).

The accepted definition of critical thinking chosen for this study incorporates the higher level cognitive functions surrounding the nurse's knowledge, prior experience, judgment, reasoning, and analysis in the provision of effective individualized nursing care. These functions are incorporated in the INACSL definition of CT which is comprehensive and inclusive of facets of other CT definitions but is also specific enough to identify nursing actions for observational measurement.

Meakim et al. (2013) state that CT is:

> A disciplined process that requires validation of the data, including any assumptions that may influence thoughts and actions, and then careful reflection on the entire process while evaluating the effectiveness of what has been

> determined as the necessary action(s) to take. This process entails purposeful, goal- directed thinking and is based on scientific principles and methods (evidence) rather than assumptions or conjecture (p. S5).

This definition is selected as it reflects the use of the term in the literature reviewed and is the definition used by nursing simulation researchers and educators. This is important as it provides a clear operational definition that is current and applicable to SBCE.

Clinical judgment.

While critical thinking is a process of analyzing data and evidence to make decisions, clinical judgment is based more on the nurse's experience and individualized knowledge of each patient in making decisions (Tanner, 2006). Tanner draws key assumptions about CJ from her 2006 literature review of nearly 200 studies on clinical judgment: (a) CJ is more influenced by the nurse's experience and attributes that they bring to the situation rather than the objective data (health care information) on the patient situation; (b) comprehensive CJ rests to a certain degree on knowing the patient and his or her typical responses, determined by engagement with the patient and his or her situation; (c) CJ is influenced by the nursing care unit and the context of the situation; (d) nurses employ a variety of reasoning patterns; and, (e) a breakdown in CJ often leads to reflection on practice which is critical for developing clinical knowledge and improving clinical reasoning.

Tanner (2006) interprets clinical judgment to be "an interpretation about a patient's needs, concerns or health problems and the decision to take action (or not), use or modify standard approaches, or improvise new ones as deemed appropriate by the patient's response" (p. 204). She stresses that CJ is a complex process requiring different types of knowledge other than that derived from science which is generalizable and applicable in many situations. CJ is knowledge that comes from the experience of applying scientific abstractions regarding practice to the individualized knowledge of a particular patient and requires a flexible and nuanced ability to recognize salient facets of an ambiguous clinical situation, interpret their meanings and respond appropriately (Tanner).

Tanner's 2006 review of the published descriptive research literature on clinical judgment in nursing resulted in the creation of her model of clinical judgment (CJ) based on this data. Tanner's CJ model includes four phases of the process of clinical judgment: noticing (grasping the situation at hand); interpreting (developing an understanding of the situation); responding (deciding on a course of action); and reflecting (attending to patient's response while taking action). Victor-Chmil and Larew (2013) assert that Tanner's key concepts of CJ set it apart from other definitions of CJ and from the concepts of critical thinking and clinical reasoning. This is accomplished through her focus on CJ as processes of cognitive and psychomotor actions and the affective process of the caregiver.

In 2007 Tanner's CJ model formed the conceptual basis for Lasater's qualitative study on the development of clinical judgment using high fidelity simulated clinical experiences. Lasater defines clinical judgment as the thinking and evaluative processes that focus on the nurse's response to a patient's multi layered problem. She further argues that CJ is highly contextual and is "the deliberate conscious decision-making characteristic of competent performance" (p. 270). Data (from simulation scenarios, debriefings and focus groups) from Lasater's (2007) study are the basis for her work on operationalizing Tanner's four phases of the CJ process. Lasater (2010) summarizes clinical judgment as "the marriage of knowledge and practical experience" (p. 87).

Alternatively, Bambini et al. (2009) in their pretest post-test design study with undergraduate nursing students with a high fidelity simulated experience, identify CJ as the ability to prioritize, identify abnormal findings and know how and when to intervene. The INACSL standard terminology definition of CJ (Meakim et al., 2013) is based on the above mentioned studies:

> It is the art of making a series of decisions to determine whether to take action based on various types of knowledge; the individual recognizes changes and salient aspects in the clinical situation, interprets their meaning, responds appropriately, and reflects on the effectiveness of the intervention. Clinical judgment is influenced by the individual's previous experiences, problem-solving, critical thinking, and clinical-reasoning abilities (p. S4).

This INACSL definition is selected for this study because it is based on formative studies (Tanner, 2006; Lasater, 2007 & 2011) identifying aspects of CJ thus providing key terms for the search and retrieval of studies applicable to simulation research around competence outcome measurement. These key CJ terms are used to develop the database search terms to capture relevant studies.

Clinical reasoning.

The term 'clinical reasoning' is often used interchangeably in the nursing literature with the terms critical thinking, problem solving and decision making (Tanner, 2006). Levett-Jones et al., (2010) also note that CR includes clinical decision-making, as did Tanner. Decision-making needs to be distinguished from CJ. The difference between CJ and clinical decision-making is clarified simply as: clinical judgment is deciding what is wrong with a patient while clinical decision-making is deciding what to do in the situation.

Lapkin, Levett-Jones, Bellchambers, and Fernandez (2010) and Tanner (2006) both define CR as a process. Tanner describes CR as the process by which nurses make clinical judgments as they select from alternatives, weigh evidence, and use intuition and pattern recognition. Lapkin et al. elaborate on the CR process as " a logical process by which nurses (and other clinicians) collect cues, process the information, come to an understanding of a patient problem or situation, plan and implement interventions, evaluate outcomes, and reflect on and learn from the process" (p. e209). Simply, CR is a non-linear process that can be conceptualized as a complex cycle of linked clinical encounters where evaluation and reflection are important elements throughout (Levett-Jones et al.). Lapkin et al. note that developing the skills of CR enhances the nurse's ability to build on past experiences and knowledge in unfamiliar circumstances. Both CR and CJ are cognitive processes that are supported by intuition and knowledge acquired through professional experience (Banning, 2008). It is this cognitive process of thinking about healthcare information that informs decisions pertinent to patient management (Simmons, 2010). CR is considered an essential component of competence (Banning).

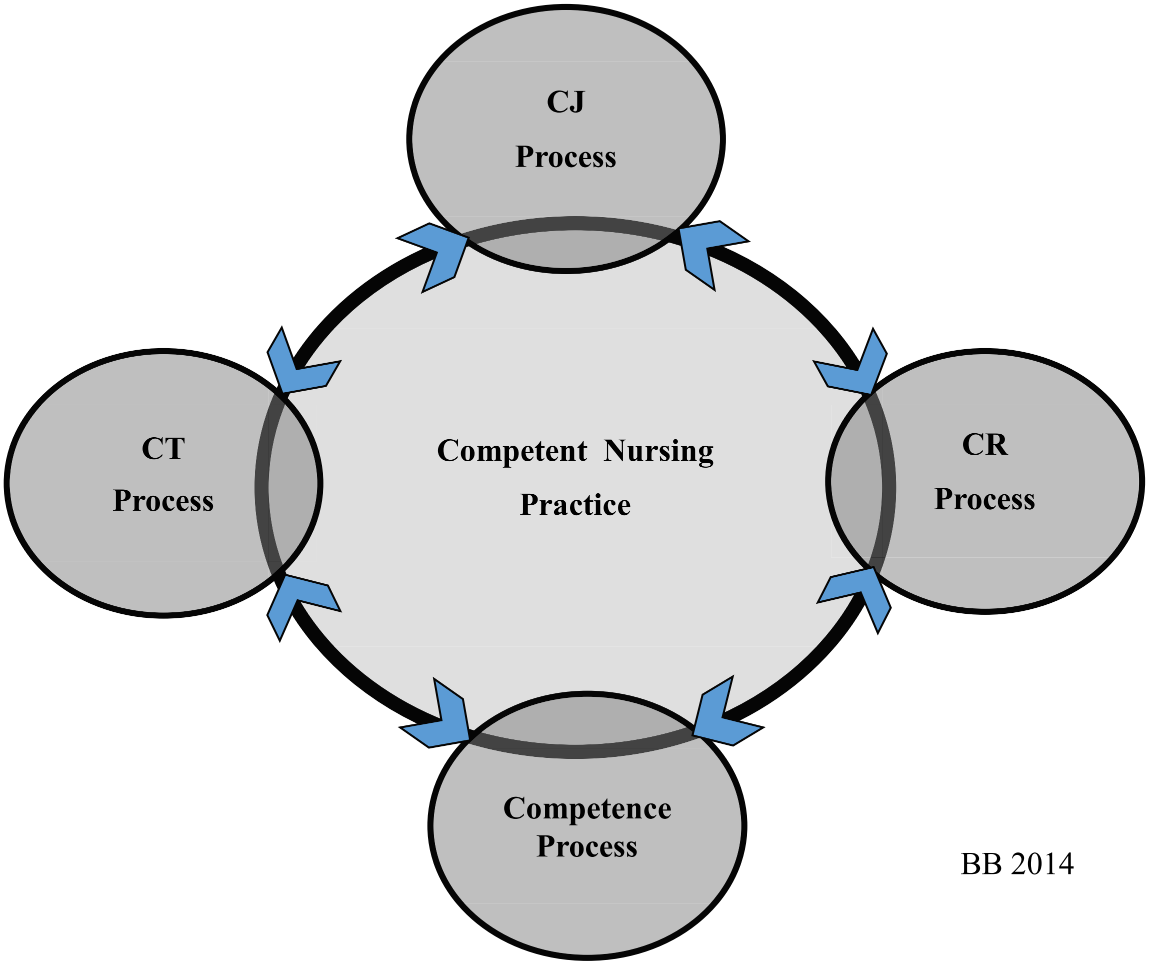
While CR is considered an essential component of competence it isn't clear at what practice level it is displayed. One view is that it is the hallmark of the expert nurse while the alternate view is that nurses at all levels employ CR skills in their decision making (Banning, 2008). Furthermore, numerous variables affect the nurse's use of the CR process including life experience, cognitive ability, maturity, and skill level in practice (Simmons, 2010). These variables explain the novice to expert levels of CR abilities.

Meakim et al. (2013) suggest the following definition of clinical reasoning in the INACSL standard 1: “the ability to gather and comprehend data while recalling knowledge, skills, (technical and non-technical), and attitudes about a situation as it unfolds. After analysis, information is put together into a meaningful whole when applying the information to new situations” (p. S4). Although definitions selected for the previous terms in this study are standard terminologies from Meakim et al., the CR definition will use a different source. The definition of CR by Meakim et al. lacks the specificity required for an operational definition for the proposed study.

This study uses the CR definition presented by Lapkin et al. (2010) as it clearly defines the process of CR into various actions that may lend themselves to measurement in SBCE. It is: “ a logical process by which nurses (and other clinicians) collect cues, process the information, come to an understanding of a patient problem or situation, plan and implement interventions, evaluate outcomes, and reflect on and learn from the process” (p. e209).

The discussion of competence, CJ, CR, and CT concepts and definitions reveal the inter-relatedness of these processes in the achievement of competent practice which is depicted in Diagram 1. The central outcome of competent nursing practice is surrounded and intersected by the circle of processes leading to its continual development. Each one of the outside circles represents one of the four processes involved in competent practice: CT, CJ, CR, and competence and each contributes its own critical component to competent practice, but is also part of a continuous cyclical process. The outer circles are all connected to each other in this cycle with backward and forward flow as well as flow into the inner circle at their point of overlap. This is the process whereby the nurse cycles through the skills related to any or all of the processes necessary for competent practice. This cycle is repeated as often as necessary in any direction of flow between the outer circles and the final outcome of competent practice.

Diagram 1 Key Processes of Competent Nursing Practice



The central outcome of competent nursing practice is surrounded and intersected by the circle of processes leading to its continual development.

In selecting the terms competence, critical thinking, clinical judgment, and clinical reasoning for inclusion as operational definitions for this systematic review, this study aims to capture all relevant studies using these terms where competence is being measured in SBCE.

SBCE and Clinical Competence Evidence

SBCE and the degree to which these experiences lead to the attainment of clinical competencies, has been studied by nursing scholars with varied methods and results. Jeffries (2007) states that SBCE lends itself to developing and evaluating five outcomes: knowledge, skill performance, learner satisfaction, CT, and, self- confidence. In fact, literature shows that SBCE is used more for evaluation purposes to measure student competencies and CT skills (Radhakrishnan et al. 2007; Todd et al. 2008). Several systematic reviews offer insight on the progress of research on SBCE use to attain competence- related outcomes.

Lapkin et al. (2010) conducted a systematic review on the effectiveness of human patient simulators in teaching CR skills to undergraduate nursing students and found inconclusive results. However, there was evidence that it significantly improved three outcomes integral to CR: knowledge acquisition, CT and the ability to identify deteriorating patient status. Limitations common to the reviewed studies are: small sample sizes; use of convenience sampling limiting generalizability; limited analysis of results; and, missing data regarding research methods. The authors note that this is an indication of the enormous challenges inherent in evaluating metacognitive processes such as CR and complex techniques such as SBCE.

Yuan, Williams, and Fang (2012) in their systematic review (2000-2011) on the contribution of high fidelity simulation to students' competence and confidence include a meta-analysis on 23 studies (18 English, six Chinese) revealing mixed contributions of SBCE to confidence and competency outcomes. They report similar limitations of included studies: a lack of high quality randomized control trials, small sample sizes, failure to report validity and inter-rater reliability of instruments, and use of a variety of measurement tools designed for traditional clinical assessments rather than specific to SBCE. Using tools out of the context for which they were meant may lead to difficulty in controlling for the variance in evaluation methods thus creating a potential bias in quantifying results (Yuan et al.).

A major challenge in this area is a lack of formal measurement tools to evaluate SBCE outcomes, and the use of self-report instruments which may lead to biased and inaccurate results (due to poor recall or inherent bias) in comparison with observation methods. They recommend the development of standardized objective measurement instruments specifically designed for SBCE and student evaluation. Even with the development of these instruments there remains the issue regarding the level of performance that indicates competence. More specifically, one may question whether competence can be assessed by focusing on individual competencies in light of the interaction between competencies.

A related concern is raised by Adamson et al. (2013) in their updated literature review of recent simulation evaluation instruments and SBCE outcome measurement. They argue that most simulation evaluation instruments focus on low-level learner outcomes such as reaction and cognitive learning rather than the higher levels which are participant's behaviours and patient outcomes. They state that low levels of evaluation may not reflect the effects of SBCE on the most important stakeholders in health care education: the patient. Furthermore, the research literature on simulation in education continues to be descriptive in design rather than empirical, and those that are empirical have questionable rigour. These concerns are also echoed by Gunberg Ross (2012) who completed a systematic review of quantitative studies related to use of simulation in the acquisition of psychomotor skills.

Despite these concerns some studies report significant results regarding competence outcomes and SBCE (Alinier et al., 2004 & 2006; Frontiero & Glynn, 2012; Lapkin et al., 2010; Radhakrishnan et al., 2007). Alinier et al. (2004) carried out a pretest, post-test design with Objective Structured Clinical Examinations (OSCEs) to determine the effects of realistic simulation experiences with human patient simulators on nursing students' competence and confidence. Nursing students were randomly assigned to control or experimental groups. After the first OSCE (which established baseline measures of clinical and communication skills for both groups) and before the second OSCE, the experimental group were given two simulation sessions. The simulation was designed to give students a clinical experience in a safe environment while avoiding specific preparation for the second OSCE.

Results showed a significant ($p < 0.05$) difference in OSCE scores between the experimental and control groups indicating that the simulation sessions had a positive effect on the students' skills and knowledge. Limitations noted were: the students volunteered to participate, many were mature students and students in both groups may have gained practical clinical experience in their normal clinical rotations during the study time. Strengths include: (a) overall design and that the content was piloted with nursing students to test the different aspects of the study; (b) the validity and authenticity of SBCE scenarios were assessed by a panel of experts and required amendments were made; (c) a second and final pilot was conducted to retest the tool; (d) use of control and experimental groups; and (e) the two OSCE sessions were identical in content (Alinier et al., 2004).

In 2006 Alinier et al. reported results from their 2004 study, this time with a larger population (N=99) and different clinical skills while still using the OSCE pretest post-test design. This time the experimental group exposed to SBCE showed a significant improvement ($p< 0.001$) in their performance compared to the control group demonstrating that SBCE has a positive impact on student performance. The reported limitation was that the two OSCEs and the simulation experience were not part of the students' curriculum so they participated on their own time, thereby limiting the number of participants. As well, the students may have gained some clinical experience from their normal study program which was not accounted for as a variable.

Radhakrishnan et al. (2007) measured specific clinical performance skills (safety, basic assessment, prioritization, problem-focused assessment, interventions, delegation, and communication) in their quasi- experimental pilot design. They evaluated the effect of human patient simulator experience on diploma nursing students' clinical skill performance. Students were randomly assigned to either the control or experimental groups (N=6) whereby the experimental group experienced the simulation practice before an end of term SBCE evaluation using a faculty developed Clinical Simulation Evaluation Tool (CSET). Results showed a significant ($p < 0.05$) difference in scores between the groups in the areas of safety and basic assessment skills only. Study strengths included random assignment, an objective instrument that rated the presence or absence of an action rather than a subjective faculty evaluation and post-test administration by an independent evaluator. Limitations noted were: the small and homogenous sample size limited generalizability, no alternate experience for the control group and no pretest baseline simulation evaluation.

Frontiero and Glynn (2012) also used a simulation evaluation instrument (developed by Todd et al., 2008) to assess student performance related to clinical judgment. Their descriptive correlational study had a convenience sample of ten senior level undergraduate nursing students experiencing a SBCE which compared the student caring for two versus four patients. The mean scores in both the two and four patient assignment scenarios were higher for assessment, CT and communication outcomes while the mean score for technical skill performance was lower in both patient-care scenarios. The authors report that these results provide support for using a simulation evaluation instrument to measure student's performance and critical thinking skills in a summative evaluation. Limitations of this study were the small sample size, one geographical location and sample homogeneity.

Several studies report non-significant and/or inconclusive findings regarding SBCE and CR skills (Jensen, 2013; Kuiper et al., 2008) and CT skills (Ravert, 2008). Jensen's descriptive study uses the Lasater Clinical Judgment Rubric (LCJR) to evaluate baccalaureate (BN) and associate (AD) degree nursing students' CR skills during SBCE to compare faculty ratings with student self-assessment ratings with both using the LCJR. Results indicate that the majority of students (63%) demonstrate adequate levels of CR in their first simulation experience and that the LCJR was able to distinguish significant differences between the total mean score for the AD and BN groups, supporting greater CR skill attainment in the BS group. However, due to the small sample size statistically significant results were not possible and significant relationships between student and faculty ratings were minimal. A major limitation of the study was the lack of inter-rater reliability of faculty using the LCJR.

Ravert (2008) in her three group pretest post-test research design, assessed CT between undergraduate BSN nursing students (N= 28) exposed to different teaching strategies (i.e. high fidelity patient simulator with enrichment sessions and non-simulator with enrichment sessions) and the control group with standard curricula; only group one had exposure to SBCE. CT was then assessed via the California Critical Thinking Disposition Inventory (Facione, Facione, & Sanchez, 1994) and the associated skills test. All three groups demonstrated moderate to large effect size improvements in CT, however there was no significant difference ($p < 0.05$) among the groups. Limitations included homogeneity within a small sample size, participant characteristics, simulator and instrument issues.

Kuiper et al. (2008) also compared SBCE with standard curricula outcomes, this time with authentic clinical experiences instead of regular classroom study. They projected outcomes of skill competency, competence, and self-efficacy in clinical practice. The purposive sample of senior undergraduate nursing students (N=44) was evaluated using the Outcome Present State-Test Model (OPT) rating tool. A comparison of the two groups indicated no significant difference ($p=.504$) between their mean scores on clinical reasoning and the paired sample t-test revealed no significant difference between the authentic clinical experience and their high-fidelity simulation experience.

Limitations included: small sample size, participant characteristics, the descriptive design, and uncertainty regarding the type of authentic clinical experiences that were compared to SBCE.

Two studies reported narrative only findings regarding SBCE and competence: Childs and Sepples (2006) and Lasater (2007b). Childs and Sepples participated in a three year multi-site study that examined the development and the implementation process of simulation in nursing education. They note that nursing students attained knowledge, acquired CT and psychomotor skills and developed confidence in their attributes through various active learning strategies including SBCE and transferred these skills into the clinical setting.

Lasater's (2007b) qualitative study focused on students' experiences using SBCE as part of their regular curriculum, focusing specifically on their development of CJ following high fidelity simulation. Junior students (N=48) participated in one simulation session a week in lieu of their regular clinical site practicum and at end of term reported their perceptions of SBCE versus regular clinical practicum via focus group design. Students reported the SBCE required them to think critically about interventions and anticipate occurrences. Lasater concludes that SBCE has the potential to support and affect development of CJ in nursing students and that SBCE could be an adjunct to clinical practice. Limitations included a small sample size and limited cultural and ethnic diversity.

Competence-Based Outcomes

In order to assess student competence following SBCE it is necessary to identify the measurable outcomes. Kardong-Edgren et al. (2010) warn that if outcome expectations are not clearly defined any instrument developed to measure student performance will fail. Each reviewed SBCE study provided outcomes in either: broad outcomes (e.g. safety, communication); by the terms CT, CJ, CR, or competence; or by more specific outcome behaviours (e.g. identifies deteriorating patient condition). The following section highlights some commonly used outcomes in SBCE research to assess nursing competence, CT, CJ, and CR. Some outcomes are listed under more than one of the competent practice processes which is to be expected due to the inter-related nature of CT, CJ, CR, and competence processes.

This systematic review only includes outcomes considered measurable by observable student behaviours. Observable is defined as “that which is capable of being observed” or “capable of being seen” (Miriam Webster Dictionary online). Observable outcome behaviours for this study are indicated with an asterisk in Table 2. Each column is independent of the other columns and lists the specific outcomes identified for each process as it was presented by the researchers of the cited studies.

Table 1
SBCE Literature Outcomes Comparison by Process

***indicates observable behaviour**

Competence Process Outcomes	**CT Process Outcomes**	**CJ Process outcomes**	**CR Process outcomes**
Knowledge gained	Assessment skills*	Noticing: focused observation* recognizing deviations from expected patterns* information seeking*	Knowledge acquisition
Communication*	Communication*	Interpreting: prioritizing data making sense of data	CT (this was a reported outcome of CR in Lapkin et al., 2010) *(See 1. Below table)*
Skill performance*	Technical skills*	Responding: calm, competent manner* clear communication* well-planned interventions/flexibility	Clinical skill performance*
			Identify deteriorating patient status*
Confidence	Disposition: truth-seeking systematic analysis inquisitiveness	Reflecting: being skillful evaluation self-analysis commitment to improvement	Describe patient situation*
Recognizing patient deviations*	Skills: analysis inference interpretation evaluation explanation*	CR (Levett-Jones et al., state this as an outcome of CJ.) *(See 2. Below table)*	Collect new patient information*

Table 1
SBCE Literature Outcomes Continued

Competence Outcomes	CT Outcomes	CJ outcomes	CR outcomes
Information seeking*		Safety|*	Review information*
Technical skills*		Interventions*	Relate information
Data prioritization		Delegation*	Recall knowledge
			Interpret information and make inferences
			Discriminate between relevant and irrelevant info
			Match & predict information
			Synthesize info to identify a problem
			Establish goals
			Choose a course of action*
			Evaluate
(Alinier et al.,2004; Bambini et al., 2009; Blum, Borglund, & Parcells, 2010; Frontiero & Glynn, 2013; Jeffries, 2005 & 2007; Radhakrishnan et al., 2007; Ravert, 2008)	(Brunt, 2005a; Frontiero & Glynn, 2012; Ravert, 2008).	(Bambini et al., 2009; Jensen, 2013; Radhakrishnan et al, 2007; Lasater, 2010; Tanner, 2006).	(Lapkin et al., 2010; Levett-Jones et al., 2010).

Table 2 SBCE literature outcomes. Each column is independent of the other columns and lists the outcomes specified for the cited studies.

1. CR is a non-linear process wherein CT is closely linked (Smith, 2012).
2. CJ is deciding what is wrong, while decision making is deciding what to do. Cr as a decision making skill can be an outcome of CJ (Levett-Jones et al., 2010)

Measurement Tools and their Rigour

The goal of measuring SBCE outcomes is to assess how well the student demonstrates learning in three domains: cognitive, affective, and psychomotor. The fourth domain of learning, the conative domain, is not mentioned in the SBCE literature to date. While the conative domain is not mentioned in the SBCE literature, it is likely that some facets of this domain are being evaluated as Huitt and Cain (2005) note that the conative domain is so intertwined with the cognitive, affective and psychomotor domains that it is rarely assessed alone and is difficult to do so. Reeves (2006) proposes that assessment of learning in higher education should encompass all four domains. Reeves states that regardless of field or discipline, there are meta-outcomes that cut across the four domains. Included in the scope of meta-outcomes are: accessing and using information; applying rules and procedures to structured and unstructured problems; thinking critically; making sound judgments; problem solving; being committed to life-long learning; proactively seeking to extend knowledge in one's discipline, and exhibiting ethical behaviour.

To measure outcomes it is essential to have valid and reliable instruments to measure performance in SBCE which are designed to assess the knowledge, values and abilities essential to competent practice, learned via the affective, cognitive and psychomotor learning domains (Jeffries, 2005). Conative outcomes, when measured alone, are most frequently measured by self-report tools (Huitt & Cain, 2005).

Learning Domains

The psychomotor learning domain includes the acquisition of technical skills which usually incorporate the affective and cognitive aspects of learning (Kardong-

Edgren et al., 2010). This is part of motor skills acquisition, as skills follow the cognitive processes of thinking and recall and often include emotional aspects of values, attitudes and beliefs related to the patient care situation. The affective domain learning of values, attitudes and beliefs, that are consistent with the standards of professional practice, may be demonstrated in SBCE as students translate them into their SBCE patient care.

In the SBCE literature, cognitive domain learning is demonstrated by cognitive actions (understanding, thinking and learning) as applied through the performance of psychomotor actions (Reeves, 2006). The conative domain focuses on the act of striving to perform at the highest levels and includes the willingness, volition, and ethics to act (Reeves). Conation is the connection of knowledge and affect to behaviour and is the intentional, deliberate, goal-oriented component of motivation (Huitt & Cain, 2005). Masui and De Corte (2005) note that conation is subdivided into motivation and volition, where motivation facilitates the formation and promotion of decisions while volition facilitates their implementation. Conative learning is particularly important in that it is closely associated with the concepts of agency, self-direction and self-regulation (Huitt & Cain), all qualities that are important to competent nursing practice.

Ideally, SBCE measurement instruments should include some measures for each learning domain as, in most instances, behaviours from several domains occur at once. Measurement in only one domain will not accurately evaluate overall performance (Kardong-Edgren et al., 2010). Comprehensive skill set SBCEs incorporate a broad range of learning domains and outcomes for evaluation, therefore they require instruments designed to measure the domains of learning. Findings from the Kardong-Edgren et al. review note that several of the tools (especially those in the cognitive group) would work in multiple domains. Cognitive domain instruments are the most comprehensive and have the best chance of measuring all three domains of learning (Kardong-Edgren et al.). Psychomotor tools are not often found in the SBCE literature perhaps because many of these tools are based on simple task trainer models not SBCEs. Alternately, affective domain instruments tend to measure participant self- confidence and satisfaction with the SBCE which have limitations due to the subjective nature of the data collected. Conative domain tools tend to measure conation with one aspect of conation and an aspect of cognition, affect, or behaviour. Most commonly measured outcomes are self-regulation, motivation, volition, and self-directed learning strategies using self- report tools (Huitt & Cain, 2005). This study focuses on only the cognitive and psychomotor domain instruments as affective and conative domain instruments are not designed for observer assessment.

Validity and Reliability of Instruments

Instruments not only need to measure the appropriate domains of learning but must do so by meeting standards of rigour to minimize bias and control for extraneous variables. Validity and reliability of measurement instruments are critical aspects of instrument design and implementation. Validity is the degree to which an instrument measures what it intends to measure. Expected instrument validity includes content validity (as a minimal expectation) and construct validity (Kardong-Edgren et al., 2010; Polit & Beck, 2008).

Content validity.

Content validity is relevant for both cognitive and affective measurement tools and regards the appropriateness of the sample items and comprehensiveness of the measurement (Polit & Beck, 2008). Basically it is concerned with how representative the selected questions or behaviours are of the whole concept under study. For example, in measuring competence if we only include the outcome behaviours of CJ, are they representative of the whole concept of competence or must other areas of this concept be represented as well? Determining content validity is a judgment decision as there are no objective methods to ensure adequate content coverage (Polit & Beck). Currently the best method to determine content validity is instrument review and evaluation by a panel of substantive content experts.

Construct validity.

Construct validity seeks to establish that the desired action adequately represents the concept under evaluation and answers the questions: what is this instrument actually measuring and, does it measure this concept adequately? Construct validity can be determined by administering the instrument to several groups who are hypothesized to be different in the concept under study and by comparing their results to see if the relationship holds true (Polit & Beck, 2008). Similarly, construct validity can be testing hypothesis relationships based on theory. For example, if one instrument measures construct A and another instrument measures construct B (where both constructs are hypothesized to have a positive relationship with each other) and the test results show the A and B correlate positively, then it is inferred that A and B are valid measures of both constructs. While this is not proof of construct validity, it does provide evidence towards that conclusion (Polit & Beck).

Criterion validity.

Criterion validity measures how well one or several items in the tool predict success on all other measures. This is considered a desired measure of validity, but is often difficult for new tool developers to establish (Kardong-Edgren et al., 2010). This is due to the difficulty in establishing a concrete, reliable criterion against which to measure the concept or outcome. Criterion validity is more easily established when a scale clearly measures the criterion (e.g. desire to lose weight measured by weight loss as measured on a calibrated scale) versus a less obvious relationship (e.g. level of professionalism by counting the number of professional meetings attended). The more abstract the concept, the less suitable it is to criterion related validity, as the selected criterion may or may not fully capture the concept under study (Polit & Beck, 2008). SBCE evaluation tools are often developed by faculty new to the area of instrument development, therefore the tools themselves are either new or recently established and require further testing to establish validity. Adamson et al. (2013) note that it is very difficult to establish validity of performance-based evaluation tools because they are often subject to the perception, knowledge, experience, and training of the evaluators.

Reliability.

Reliability in quantitative instrumentation is a key criterion in assessing its adequacy and quality (Polit & Beck, 2008). Reliability is the consistency with which the instrument measures its' intended attribute in different and repeated situations. A measure is said to be reliable when it maximizes the true score component and minimizes the error component. Reliability is associated with the measures' stability, consistency and dependability. Stability concerns the extent to which similar results are obtained on two or more separate occasions and this is test-retest reliability. It is established objectively by computing the reliability co-efficient which expresses the magnitude of the test's reliability. Statistically this is expressed by a correlation coefficient which quantitatively describes the direction and magnitude of the relationship between two variables. The possible values for the correlation coefficient range from -1.00 to +1.00, with 1.00 expressing a perfect relationship (Polit & Beck).

A second measure of reliability is internal consistency (the extent to which all the instruments' items measure the same attribute) and is often used when scales and test tools involve summing of item scores (Polit & Beck, 2008). A commonly used method for establishing internal consistency is the statistical measure of Cronbach's (or coefficient) alpha which conveys the extent of consistency. The normal range of values lie between .00 and +1.00 where higher values express a higher level of internal consistency.

The third and critical measure of instrument reliability is equivalence reliability which concerns inter-rater reliability, measuring the degree to which two observers agree in their scoring (observations) on an instrument. A high level of consistency in scores leads to the assumption that measurement errors were minimized. Inter-rater reliability can be established by consensus, consistency and by measurement methods (Polit & Beck, 2008). In observational coding, consensus measures involve having two or more trained observers watching an event at the same time and independently scoring according to the instruments' instructions. This data is then computed to express an index of agreement between the observers. Cohen's kappa is a commonly used statistic and multirater kappa is used when there are more than two raters. A value of .60 is minimally acceptable and values of .75 or higher are considered excellent. As well, an intraclass correlation coefficient can be used to demonstrate the strength of the relationship between the raters' scores. Inter-rater reliability is described as being somewhat difficult to establish in SBCE as the SBCE often portrays rapidly changing situations in which participant behaviours also change quickly making it tricky for observers to capture and evaluate outcomes (Kardong-Edgren et al., 2010).

Measurement Tools

Adamson et al. (2013) stress that although an instrument may score well in validity and reliability the educator must select the appropriate tool based on their intended SBCE activity, student population sample, and by the number of raters and their level of experience. This includes taking note of whether or not the instrument is being used for a measurement purpose different from the originally intended purpose. For example, if the instrument was designed to measure CJ outcomes with undergraduate students following SBCE but the next researcher intended to use it to measure skill competency in assessments with graduate students and actual patients, then the test may not be valid or reliable in these new circumstances. If the instrument is to be used for a new purpose it is advisable for the researcher or educator to check with the instrument developers regarding their processes of instrument validation and reliability with the associated statistics. In order to validate the instrument for the new purpose it should be tested via pilot projects and content expert reviews (Adamson et al.).

The review of the literature reveals five types of measurement instruments associated with SBCE and competence outcomes: OSCE, rubrics, simulation evaluation tools, checklists, and paper and pencil CT tests. Rubrics and simulation evaluation tools have been used repeatedly in SBCE research. Instruments that measure the cognitive domain include: the Sweeny-Clark Simulation Performance Evaluation Tool (Clark, 2006 Clark Tool©), the Lasater Clinical Judgment Rubric (LCJR) (Lasater, 2007), the Clinical Simulation Evaluation Tool (CSET) (Radhakrishnan et al., 2007), and the Creighton Simulation Evaluation Instrument (C-SEI™) (Todd et al., 2008). A more recent simulation evaluation tool is The Seattle University Simulation Evaluation Tool© (Mikasa, Cicero, & Adamson, 2013). A checklist method evaluation tool was employed by Wolf et al. (2011) while critical thinking assessment tools (California Critical Thinking Test (CCTST) (Facione & Facione, 1998) and the California Critical Thinking Disposition Inventory (CCTDI) (Facione, et al., 1994) were employed by Ravert (2008).

OSCE.

An OSCE (Objective Structured Clinical Examination) provides an observational method of student competence as it is a set of performance-based scenarios in which the student is observed demonstrating clinical behaviours and interventions. OSCE's have been used mainly to evaluate cognitive learning in performance related outcomes. Students rotate through the various stations and are assessed by observer testers using checklists and rating tools. A SBCE may be the only station of the OSCE or it may be only one of the various techniques used to test student abilities and provide outcomes for measurement. OSCEs require four areas for consideration when evaluating reliability and validity: measuring context-reliant competence, measuring competence versus performance, measuring professional behaviour, and measuring integration of skills

(Adamson et al., 2013). Inter-rater reliability of testers in evaluating student performance is also important.

LCJR tool.

The Lasater Clinical Judgment Rubric (LCJR) was designed to evaluate CJ concepts presented in Tanner's (2006) work. The LCJR: assesses CJ through clearly specified action outcomes; has also been used to assess debriefing post SBCE and technical skill performance; and has had extensive reliability and validity findings from a range of studies to assess its psychometric properties (Adamson et al., 2013).

The Clark© tool.

The Clark Tool© has a framework based on Bloom's taxonomy and Benner's novice to expert levels of experience (Clark, 2006) and was originally used in an obstetrical trauma simulation. Because of its strong framework it has allowed for modification to other scenarios for both undergraduate and graduate student participants. Kardong-Edgren et al. (2010) comment that inter-rater reliability is easily established with this tool and reported a Cronbach's alpha of > .86 for internal consistency (Clark, 2006).

CSET tool.

The Clinical Simulation Evaluation Tool (CSET) (Radhakrishnan et al., 2007) was designed to evaluate student competencies in safety, basic assessment, problem-focused assessment, prioritization, interventions, delegations, and communication in SBCE. It is reported as being one of the few instruments coming closest to evaluating three of the learning domains simultaneously (cognitive, affective and psychomotor) (Kardong- Edgren et al., 2010) and has been modified to suit diverse performance evaluation needs (simulation and standardized patient scenarios). Grant, Moss, Epps, and Watts (2010) report inter-rater reliability findings from Fleiss's kappa coefficients at .71 to .94 indicating a percentage agreement from 85% to 97% among the five data collectors. This literature review did not reveal any other reports of validity or reliability for the CSET.

C-SEI™ tool.

The Creighton Simulation Evaluation Instrument (C-SEI™) (Todd et al., 2008) was based on the American Association of Colleges of Nursing's core competencies for new graduates providing a group grade for collaboration in SBCE (Kardong-Edgren et al., 2010). Group assessments in SBCE are often necessary due to large class sizes and clinical group size in nursing and with the emerging trend of SBCEs with groups of inter-professional students. The tool includes 22 behaviours in the categories of CT, communication, assessment, and technical skills. Content validity was established by identifying the key concepts of each category to be measured based on evidence from the literature and by a positive content assessment with an expert panel rating the items on questionnaire via a Likert scale. The instruments' inter-rater reliability was established by training the six evaluators in information and practice sessions on the use of the tool and by pilot testing the tool and raters' scoring with (N=72) students. The inter-rater reliability was reported in terms of simple percent of agreement as 85 to 89 per cent agreement in scoring where 80 per cent is acceptable (Todd et al. 2008).

The Seattle University Simulation Evaluation Tool©.

The Seattle University Simulation Evaluation Tool© (Mikasa et al., 2013) is an outcome-based evaluation of student performance during SBCEs measuring assessment, CT, patient care, communication, and professionalism. It is an evaluation rubric used by both faculty evaluators and by the student participants who rate their performance immediately after the SBCE and prior to the debriefing session. The tool was reviewed several times by the Seattle University Center for Excellence in Teaching by an education expert on evaluation scales prior to use. It was initially implemented with N=84 students and N=7 faculty evaluators with a SBCE. Later it was included with two other instruments in a multi-site collaboration assessing reliability of simulation evaluation tools (Adamson & Kardong-Edgren, 2012). Internal consistency of reliability was assessed with Cronbach's alpha at .97. Equivalence reliability was assessed by inter-rater reliability at 0.858 using intra-class correlation due to the fact that the investigator was comparing ratings at one point in time to the same rater's ratings at another point in time. Both the inter and intra reliability analyses were based on two-way ANOVAs with individual ratings as the units of analyses.

Checklist tool.

Wolf et al. (2011) also evaluated SBCE and student performance but used a checklist method for evaluators to rate observed behaviours. Behaviours were marked as present or absent and inter-rater reliability by others using the tool in research was 95%. Face validity was established by faculty practice experts reviewing the instrument content and appropriateness of required key behaviours drawn from nursing education literature. The authors reported that the criterion-related validity was not possible at the tool's inception as there were no instruments to use at that time to measure it, but since then the subscales have been compared and found parallel to Lasater's CJ rubric (Lasater, 2007a). Predictive validity was underway at the time of publication with a version of the tool being tested by a faculty member measuring performance of new nursing staff at the end of their orientation and after one year of clinical practice.

CCTST and CCTDI.

The critical thinking measurement tools the California Critical Thinking Test (CCTST) (Facione & Facione, 1998) and the California Critical Thinking Disposition Inventory (CCTDI) (Facione, et al., 1994) were developed by the American Philosophical Association to assess critical thinking in college students but are not specific to nursing students. These instruments are completed by the participants not the evaluators. The CCDTI has an internal consistency measured by Cronbach's alpha at 0.91 while the CCTST has an internal consistency of 0.68 to 0.80. Equivalence reliability, stability or validity were not mentioned.

Adamson et al. (2013) stress that researchers can advance simulation pedagogy by reporting psychometric measures and the steps taken to assure validation of the instrument when it is used with new populations. Validation information, along with aspiring to higher levels of evaluation outcomes, would help to indicate how SBCE affects learning, behaviours and patient outcomes. Such outcome measures are possible with SBCE because it offers evaluation of the higher cognitive functions of application, synthesis and evaluation of nursing knowledge (Kardong-Edgren et al., 2010).

This systematic review seeks to review SBCE studies employing any one of the above mentioned instruments except for the CCTST and the CCTDI because these instruments are not observational measures (as they are completed by the participants not the evaluators).

Systematic Review Overview

The systematic review method is a structured scientific method of scholarly inquiry to assemble, critically appraise, and synthesize research evidence from pertinent studies regarding a specific clinical problem or topic for development (Windle, 2010). The systematic review researcher accepts that there is a hierarchy of evidence derived from the selected studies whose design is explicit and rigorous. Houde (2009) identified four key elements of a systematic review: it uses transparent and explicit methods; the research follows standard stages; it is accountable, replicable and able to be updated; and, there is a requirement of user involvement.

The researcher prepares a protocol for the review identifying the specific search strategy which is directed by reproducible criteria which limit the risk of bias and random error in the study. The search is comprehensive and includes electronic databases as well as hand searching, searching references described in the selected studies, contacts with researchers and the use of unpublished material (grey literature).

An appraisal tool is identified prior to the data search which includes the explicit inclusion and exclusion criteria against which the selected studies are then critically appraised in an objective manner (Marshall & Sykes, 2011). The quality appraisal tool guides the extraction of specific methodological aspects of primary studies in order to evaluate the overall quality of the study. While it is recommended that more than one reviewer carry out the quality assessment to help minimize bias and error and allow for inter-rater reliability to be assessed (Webb & Roe, 2007), it does not preclude one reviewer from conducting a systematic review, but this fact must be accounted for in the interpretation of findings (Marshall & Sykes). Once appraised for inclusion or exclusion, the selected study results are analyzed for credibility to evaluate the strength of their evidence (Holopainen et al., 2008).

The researcher documents the study characteristics, quality and results. Finally, the material of the primary studies are organized, categorized (by theme or study design), and combined via statistical and narrative methods. The systematic review reports the search strategy and results, identifies the number of articles retrieved and number of articles rejected by criteria reason. The results are presented as conclusions by analysis of results or synthesis of results (Bettany-Saltikov, 2010). If primary studies cannot be combined statistically a narrative analysis is considered acceptable (Whittemore & Knafl, 2005).

Systematic Review Models

Cochrane Collaboration.

Several models of systematic reviews exist, each providing guidance for SR process steps, appraisal tools and focus on a particular research design for the primary studies to be reviewed, while still meeting the principles of a SR. A systematic review can be carried out with the support, publication and dissemination of review results via a global network of researchers. The Cochrane Collaboration and the Joanna Briggs Institute are two such organizations assisting reviewers**.** The Cochrane Collaboration systematic reviews are well recognized as the highest standard in evidence based health care and are focused primarily on health care interventions and health policy. A Cochrane review is a scientific investigation of the best evidence using pre planned methods and reviews studies of randomized controlled trials (RCTs), and clinical controlled trails and on occasion, non-randomized studies.

In order to write a Cochrane review a reviewer registers the systematic review with a relevant Cochrane Review group. The groups are based on different health areas and interventions. The guidelines state that it is essential that more than one reviewer carry out the review in order to ensure that tasks such as selection of studies for eligibility and data extraction have at least two independent reviewers, thus increasing the likelihood that errors are detected. Once the topic is accepted by a group the review title must be accepted and registered, then the authors submit the review protocol to the Cochrane group. The protocol may go through several iterations before being accepted. The purpose of publishing the protocols for Cochrane reviews (in the *Cochrane Database of Systematic Reviews)* prior to publication of the completed review is to reduce the impact of authors' biases, to promote transparency of methods and processes, to reduce the potential for duplication, and to allow peer review of the planned methods.

Once the protocol is accepted the reviewer can proceed with the study. The estimated timeline for completing all tasks related to a Cochrane review is 11-12 months or longer. These tasks include training, meetings, protocol development, searching for studies, assessing citations and full-text reports of studies for eligibility, assessing the risk of bias of included studies, collecting data, pursuing missing data and unpublished studies, analysing the data, interpreting the results, and writing the review. Authors are expected to complete the review within a reasonable time frame and to keep it up-to-date once it is completed (Higgins & Green (Eds.), 2011).

Joanna Briggs Institute.

The Joanna Briggs Institute (JBI) is an international collaboration of centres and units and a leader in conducting reviews of economic, qualitative and policy research. Their goal is to aid in the improvement of worldwide healthcare outcomes by developing and promoting evidence-based resources for service provides, healthcare professionals and consumers worldwide (Joanna Briggs Institute, 2013). Their systematic review approach is broad and inclusive of diverse sources of research-based and non-research-based evidence. It promotes the inclusion of qualitative research studies (regarded as rigorously generated evidence) and other text derived from opinion, experience, and expertise, and they are acknowledged as forms of evidence when the results of research are unavailable (Pearson, Wiechula, Court, & Lockwood, 2005). Inclusion of these diverse sources helps to situate the process of evidence- based practice within a broader context (Pearson et al.). The JBI model of evidence promotes and supports the synthesis, transfer and utilization of evidence by identifying healthcare practices that are considered feasible, appropriate, meaningful, and effective.

JBI reviews are primarily focused on point of care research and evidence in nursing and other health fields as well as medicine. The critical appraisal of research studies has two major purposes in JBI: to evaluate the quality of a study by examining it for possibility of bias (the goal is to only include studies of high quality) and to determine whether the researcher has clearly indicated the population, intervention and outcomes of interest in order to synthesize the results of the studies and make recommendations for evidence based practice or policy (Houde, 2009). JBI reviews are also registered with the appropriate review centre according to review topic and require that the researcher attend training after which the reviewer can register with JBI, submit the review proposal, and access the JBI search and management databases.

The JBI model follows seven steps: the development of a rigorous proposal or protocol which provides a predetermined plan; the statement of the questions or hypotheses that will be pursued in the review; the detailing of a strategy that will be used to identify all relevant literature within an agreed time frame; the establishment of a method to assess or critically appraise the quality of each study/paper and any exclusion criteria based on quality considerations; the description of how data will be extracted from the primary research or text; and the description of a plan on the synthesis of data (Pearson et al., 2005, p 211).

Groups with Guidelines for Reporting SR Results

Groups of researchers have established guidelines for the reporting of SR results, and some provide study appraisal tools appropriate for systematic reviews of experimental design research. Some that may be appropriate for the proposed SR are: the Consolidated Standards of Reporting Trials (CONSORT) and Transparent Reporting of Evaluations of Non-Randomized Designs (TREND) and the Cochrane Collaboration Risk of Bias tool. Each group presents checklists and flow diagrams and a handbook providing guidance for use of the tools. The Newcastle Ottawa Scale (NOS) is not appropriate for this review because it focuses on case control and cohort studies, neither of which will be included in this review.

Although the proposed systematic review seeks outcome measurements which are derived from evaluators observing participants, it is not to be confused with a SR intended to gather observational research design studies where the investigator simply observes and does not apply an intervention or attempt to alter what occurs. Non-experimental designs require different SR guidelines (Polit & Beck, 2008).

Summary

The literature supports that SBCE can supplement the development and evaluation of competence-related skills despite varied limitations of study design and results regarding SBCE and competence outcomes. Simulation pedagogy will be developed when researchers assure validation of the instrument, report psychometric measures, and aspire to higher levels of evaluation in the cognitive functions of application, synthesis, and evaluation of nursing knowledge. Such outcome measures are possible with SBCE. It is stressed that educators select tools appropriate to their intended SBCE activity, student population sample, number of raters, and their experience level.

In reviewing measurement tools and their rigour it is evident that it is essential to acknowledge the four learning domains (cognitive, affective, conative, and psychomotor) encompassed in assessing the knowledge, values, and abilities necessary for competent practice. Ideally, a measurement instrument should include some measures of each domain to provide an accurate evaluation of overall performance. Only two domains however, lend themselves to observable measurement in SBCE: the cognitive and psychomotor domains. This study aims to identify the cognitive and psychomotor domains associated with each instrument reviewed.

Two main gaps in the SBCE and competence measurement literature were identified. First, there is a lack of consensus on the definition and meaning of the concepts and use of the terms competence, CT, CJ, and CR as they relate to the profession of nursing. As well, there is a lack of clarity around these four concepts, their relationship to each other, and to the outcome of competent practice. The literature synthesis for the proposed systematic review examines the literature on all four concepts and the most appropriate definition of each is selected to be the operational definitions guiding the systematic review search. Additionally, the literature provides evidence of the inter-related nature of CT, CJ, CR, and competence and identifies each as a process leading to the outcome of competent practice. Therefore, the systematic review includes all four processes as outcomes of competence thereby leading to a comprehensive search strategy aimed at being as inclusive as possible.

A second gap in the literature pertains to the lack of clarity on what competence outcomes are and how they should be measured. Many studies use the terms CT, CJ, CR, and competence as the outcome measures of competence which leads to general findings rather than clear performance behaviours as indicators of competence. This, in turn, creates a lack of direction for others wishing to use an instrument to measure skills in specific learning domains, or for more specific outcome indicators of competence. The literature reviewed does provide more specific outcome behaviours for each of the four competence processes (as listed in Table 2) and several are subsequently identified as being observable SBCE competence outcomes for this systematic review.

Expected standards of rigour for measurement instruments include content and criterion validity and reliability measures of internal consistency, equivalence, and stability. An observation from this review reveals inconsistent evidence of design rigour among the studies and varying levels of reporting within each study. The proposed systematic review aims to address this by determining which tools provide observational outcome measures that are both valid and reliable. To that end, this review sets the study selection inclusion criteria so that only studies with some level of reported validity and reliability will be included.

The systematic review design is selected for the proposed study because its purpose is to determine the effectiveness of interventions by comparing two or more interventions. This purpose fits well with a systematic review of SBCE and outcome measurement as SBCE study designs tend to have two or more interventions to compare. The systematic review methodology involves scholarly rigour in accepting only studies whose design is explicit and rigorous while adhering to a strict protocol to limit the risk of bias and random error in both the review design and in the inclusion and exclusion of studies. This protocol helps to ensure that the results include the best evidence to answer the research question. The systematic review method is the most commonly used method to review quantitative studies which is the focus of the proposed study. Several systematic reviews on SBCE have been carried out but to date none have focused on only observational evaluation of undergraduate nursing students and the related validity and reliability of the associated measurement instrument. In this way, the proposed study will add to nursing knowledge by providing evidence to help educators and researchers involved in SBCE and competence measurement, thus advancing the body of knowledge around the scholarship of teaching within a simulated clinical environment.

CHAPTER 3 METHODOLOGY

Introduction

The aim of the proposed systematic review (SR) will be to determine which observational outcome measurement (evaluation) tools provide measures that are both valid and reliable in measuring undergraduate nursing students' outcomes of nursing competence, clinical judgment, clinical reasoning and critical thinking following a high-fidelity simulation-based clinical experience.

The SR methodology was selected as a means to collate all available empirical evidence that fits pre-specified eligibility criteria to answer a specific research question. It uses explicit, systematic methods that are selected in order to minimize bias, thus providing more reliable findings from which conclusions can be drawn and decisions made (Higgins & Green, 2011). The Canadian Institutes of Health Research (CIHR) define this synthesis of knowledge as "the contextualization and integration of research findings of individual research studies within the larger body of knowledge on the topic" (Grimshaw, 2010, p. 2) and assent, that knowledge syntheses are an "efficient scientific approach to identifying and summarizing evidence that allow the generalizability and consistency of research findings to be assessed and data inconsistencies to be explored" (p.4). As the results of this SR are potentially useful to nursing researchers, educators, and stakeholders it is critical that the findings from this study be objective, comprehensive, and reliable.

Chapter three provides a brief overview of the main principles and steps that govern a SR and how they differ from a literature review. Each step of the proposed methodology will also be presented as follows: protocol development, study inclusion and exclusion criteria and methods, appraisal and data extraction methods, methods of analysis and synthesis, and presentation of findings, along with rationale and evidence to support the methods selected.

Systematic Review Definition

Polit and Beck (2008) define a systematic review as "a rigorous and systematic synthesis of research findings on a common or strongly related research question" (p 767). Whittemore and Knafl (2005) define SR more specifically, as the systematic syntheses of findings from quantitative studies. Among most authors, a SR is considered a methodical, scholarly inquiry that follows many of the same steps as those used in primary research studies and aims to provide the best evidence at the time the review was written (Grimshaw, 2010; Polit & Beck). The Canadian Institute of Health Research (CIHR) defines knowledge synthesis as

> "the contextualization and integration of research findings of individual research studies within the larger body of knowledge on the topic. A synthesis must be reproducible and transparent in its methods, using quantitative and/or qualitative methods. It could take the form of a systematic review…" (p 2) (Grimshaw, 2010).

Purpose of a Systematic Review

The purpose of a knowledge synthesis is to summarize evidence around a specific question (Grimshaw, 2010) and can be used to guide clinical, leadership, and educational decisions (Houser, 2012). The SR also generates a record of existing research, sets out what is known about a particular intervention or question, and can demonstrate knowledge gaps (Centre for Reviews and Dissemination (CRD), University of York, 2009), which can then be used to guide future research.

The purpose of the proposed SR research study is to carry out a systematic review of the literature from no specified start date to November 2014 to determine which outcome measurement tools to date provide observational outcome measures of nursing competence, clinical judgment, clinical reasoning, and critical thinking that are both valid and reliable for undergraduate nursing students following simulation-based clinical experiences. This study is based on the assumption that these outcomes can be measured by observation.

Systematic Review versus Literature Review

Systematic reviews are often mistaken for literature reviews but, in fact, are quite different in various features of their process and reporting. The literature review may focus on a single question but also may be an overview of an area of interest while a systematic review question is one that is specific and focused. A literature review does not include a protocol whereas a systematic review is based on a protocol or plan created before the search begins. Clear criteria for the inclusion or exclusion of studies are identified before the systematic review is conducted however, there are no criteria specified in a literature review. As well, literature reviews do not explicitly state the search strategy, whereas this is considered an important step in a SR, where a clearly mapped out search strategy is carried out methodically (Bettany-Saltikov, 2010).

Systematic reviews have transparent processes for selecting and evaluating research articles and for extracting data (relevant information) whereas, similar processes are not described in a literature review. Literature reviews may or may not evaluate the quality of the studies reviewed but a SR incorporates this evaluation comprehensively. One of the more important differences between the two review types is that SRs provide clear summaries of the studies based on high quality evidence while summaries from literature reviews may be based on studies of unspecified quality and may be influenced by the reviewer's beliefs, theories or needs (Bettany-Saltikov, 2010).

Principles and Characteristics of a Systematic Review

The differences between a literature review and a systematic review point to the key characteristics that are responsible for the SR's inherent quality and trustworthiness of results. Houde (2009) identified four key elements of a systematic review: the use of transparent and explicit methods; standard stages of research are followed; the review is accountable, replicable and able to be updated; and the requirement of user involvement is met. The Cochrane Handbook for Systematic Reviews of Interventions (Higgins & Green, 2011) presents the key characteristics of a systematic review as:

> …a clearly stated set of objectives with pre-defined eligibility criteria for studies; an explicit, reproducible methodology; a systematic search that attempts to identify all studies that would meet the eligibility criteria; an assessment of the validity of the findings of the included studies, for example through the assessment of risk of bias; and a systematic presentation, and synthesis, of the characteristics and findings of the included studies (p 1.2.2).

All systematic reviews share common features: they are led by a well-focused and feasible question; the SR uses explicit procedures or review protocols and methods for evaluating retrieved studies; the SR provides a transparent description of methods so that it is reproducible by another researcher who could arrive at the same conclusions; the SR acts as an efficient information management tool by reducing the volume of information on a topic; and is concerned with adding value to the research community and stakeholders.

Systematic Review Methodology Overview

Bettany-Saltikov (2010) describes the systematic methods involved in the review as the identification, selection, appraisal, and synthesis of high quality research evidence that is pertinent to the question under study. It begins with an answerable research question and the development of a detailed protocol for the study, followed by the search of the literature for relevant studies. The search strategy and the search results are saved and recorded in an electronic format providing a comprehensive list of studies that may meet the criteria for inclusion in the review and that may be appropriate for answering the research question. The reviewer includes only studies with a research design that is both explicit and rigorous (Bettany-Saltikov).

The selected studies are then critically appraised in an objective manner by the application of explicit inclusion and exclusion criteria that evaluate the overall quality of the study (Marshall & Sykes, 2011). This is determined by three phases of appraisal and extraction of information from the studies: phase one which includes selecting studies for inclusion or exclusion in the review, phase two which consists of appraising the quality of the articles, and phase three in which the data is extracted from the study. The application of these three methods in the SR aims to ensure the appropriateness of the included studies, and that the SR can be easily evaluated and replicated. All three phases are discussed with a critical review panel and/or a supervisor in the case of a graduate thesis to ensure the results are free from bias (Bettany-Saltikov, 2010).

The reviewer maintains a record of each abstract and article reviewed, along with the reason for elimination if a study was dropped from consideration. This ensures that the reviewer is using objective and defensible reasons for selecting studies for final recommendations and minimizes the potential effects of researcher bias on the outcome.

This is followed by analysis of the data, reporting of various stages of the process, presentation of the results, and discussion of the findings, which are interpreted and presented in an objective summary via narrative and/or statistical (where appropriate) methods. The researcher also reports the search strategy and results, the number of articles retrieved and number of articles rejected by criteria reason, all of which were carefully documented to enable the reader to evaluate the validity and process of the review.

Protocol Methodology for the Proposed SR

The protocol is the first step towards developing the specific search strategy, which is directed by reproducible criteria, which in turn limits the risk of bias and random error in the study (Higgins & Green, 2011). In knowledge synthesis this is referred to as a framework but includes similar steps of identifying: the objectives of the review; the eligibility criteria for studies for SR inclusion; the search; the method of recording search results; the method of appraising the quality and appropriateness of the primary studies; the specific data to be extracted from the articles; analyses of the data including narrative methods and statistical synthesis and sensitivity analyses if appropriate; syntheses of the evidence with use of tables for comparisons and analysis; and finally, preparation of the research report (Grimshaw, 2010; Higgins & Green, 2011). The protocol limits the likelihood of biased post hoc decisions in the review methods (Moher, Liberati, Tetzlaff, & Altman, 2009).

To ensure that the proposed SR meets the SR standards of reporting, the Preferred Reporting items for Systematic reviews and Meta-Analyses criteria (PRISMA) checklist and guide will be followed as closely as possible during the review. PRISMA focuses on ways in which authors can ensure the transparent and complete reporting of systematic reviews and meta-analyses (Moher et al., 2009). PRISMA guidelines consist of a checklist, a flow diagram, and the PRISMA Explanation and Elaboration document. See Appendix B.

SR question and PICOTS elements.

A comprehensive systematic search of the nursing literature will be conducted to answer this SR research question: what observational outcome measurement tools are reliable and valid in measuring students' individual competency outcomes for simulation-based clinical experiences using high fidelity patient simulators for undergraduate nursing student clinical experiences? The SR research question is framed using the PICOTS elements, which then, guide the development of the SR study eligibility criteria and subsequent search strategy (CRD, 2009; Grimshaw, 2010).

The population for the proposed SR will include studies with undergraduate nursing students in any year of their program in either an associate degree or baccalaureate degree-nursing program. The review is focused on nursing students only; medical and allied health students are excluded due to their different knowledge, skill sets and methods of achieving competence. Additionally, graduate nursing students, newly graduated nurses, and staff nurses are excluded from the study due to their advanced level of competence and experience that would create a potential confounder in the outcome under review.

The intervention under review is high fidelity SBCE inclusive of a high fidelity patient simulator and clinical scenario with a realistic environment and equipment with the study objective to develop or test a measurement instrument tool. Excluded interventions include low or medium (moderate) fidelity SBCE, on-site clinical practicum experiences, lab-based clinical experiences, and SBCEs with standardized patients.

The comparison interventions included in this review will be on-site, clinical practicums (at a healthcare agency or community site), lab-based clinical experiences, case study clinical scenarios, standardized-patient clinical scenarios, clinical scenario lecture classes, low or medium fidelity SBCE, and low or medium clinical skill simulation experiences. Excluded comparison interventions are computer programs or virtual clinical experiences, as they are not assessed by observational measures.

The SR study outcome measures for this SR are outcome measurement tools based on evaluator observation identifying CT, CJ, CR, and/or competence outcomes. Outcomes that will be excluded from this SR are: student self-measurements of confidence, satisfaction with SBCE, competence, CT, CR or CJ; all self-rating outcome measurement instruments for competence, CT, CJ or CR including the California Critical Thinking Disposition Inventory (Facione, Facione, & Sanchez, 1994) and the California Critical Thinking Test (Facione & Facione, 1998) as they provide self-assessment outcomes via the participant (student) and do not provide the objective outcome measure which is the focus of this SR.

The proposed SR will incorporate primary research studies that include: experimental design (RCTs), quasi-experimental design (non-randomized controlled trial, comparative design before-and-after study, and interrupted time series), pilot studies of experimental or quasi-experimental designs, and observational pre and post cohort studies. This researcher accepts that the types of studies included in an SR play a major role in determining the reliability and validity of the results and ideally, a SR would include only RCTs and controlled trials as they are of more robust designs (CRD, 2009). Also noted, is that experimental study designs are best suited for answering research questions of intervention effectiveness (Grimshaw, 2010). However, this SR is not focused on the effectiveness of the outcome measurement tools, rather on the validity and reliability of the tools, thus making it acceptable to include quasi-experimental and observational study designs in the SR (CRD, 2009). Furthermore, it is not anticipated that many experimental design studies will be available in the SBCE outcome measurement literature. Qualitative studies, post-test only observational studies, commentaries, and reviews will be excluded from the review as they do not provide empirical data that can be used to evaluate validity and reliability of outcome measures sought for this study.

Study eligibility criteria.

The identification of inclusion and exclusion criteria establish the sampling frame as these criteria describe the type of studies from which data will be drawn. The establishment of the focus and limits of the SR provide a clear rationale for study inclusion or exclusion criteria, creating a rigorous process and minimizing the risk of bias, as well as avoiding inclusion of irrelevant material (Webb & Roe, 2007). Grimshaw

(2010) emphasizes that poorly specified criteria may lead to insensitive search strategies resulting in a failure of the SR to identify some or all relevant studies. As well, non-specific criteria increase the workload of the researcher in reviewing irrelevant material. Inclusion criteria must also be practical to apply; if they are too detailed screening may become overly complicated and time consuming (CRD, 2009). The criteria are derived from the research question therefore providing a description of key variables such as: the research designs applicable to the question under review, the language, time frame, and geographical location. It is essential that the direct connection between the research question and the criteria for inclusion be clearly evident (Marshall & Sykes, 2011). The researcher strictly adheres to the criteria whereby studies must meet all eligibility criteria for inclusion.

This SR has a language search limit of English only or studies already translated into English (due to the time and cost involved in having non-English studies translated) albeit any relevant non-English studies will be documented and identified with language as the reason for exclusion (CRD, 2009). The SR time frame limit is from open start date November 2014. The open start date is selected to capture past as well as recent student performance measurement instruments involving high fidelity SBCE and the end date is necessary to enable thesis completion. The geographical location is not limited but the setting limit is a high-fidelity simulation lab. The eligibility criteria and study limits for this SR are detailed in Table 3.

Table 2

Eligibility Criteria

PICOTS and limits	Inclusion Criteria	Exclusion Criteria
Population	Undergraduate nursing students -Associate degree -Baccalaureate degree -in any year of their program	-Graduate nursing students -Students of any other health program -Graduate nurses -Staff nurses
Intervention	High fidelity simulation-based clinical experience includes: - high fidelity computerized patient simulator -clinical scenario with realistic environment and equipment	Low or medium (moderate) fidelity simulations: - low or medium patient simulator - unrealistic setting (e.g. classroom, basic lab) On-site clinical practicum experiences Lab-based clinical experiences SBCEs with standardized patients
Comparison intervention	-Clinical practicum on-site at a healthcare agency or community site. -Lab-based clinical experiences -Clinical scenario case study -Clinical scenario with standardized patients -Lecture classes on clinical scenarios -Low or medium fidelity simulations -Low or medium clinical skill simulation experiences.	Computer or virtual clinical experiences
Outcomes	Observational outcome measurement instruments Student outcomes of: -CT -CJ -CR -competence	Self-rating instruments capturing - competence CT, CR or CJ (including California Critical Thinking Disposition Inventory & California Critical Thinking Test)* Outcomes of: -confidence -satisfaction with SBCE *(Facione, Facione, & Sanchez, 1994) (Facione & Facione, 1998)

Table 2 Eligibility Criteria continued

PICOTS and limits	Inclusion Criteria	Exclusion Criteria
Study design types	Quantitative: Experimental: -randomized controlled trial Quasi-experimental: -non-randomized controlled trial -comparative design before-and-after study - interrupted time series Pilot study of experimental or quasi-experimental studies Observational: - pre and post cohort studies	-Qualitative studies -Quantitative post-test only -Observational studies -Case studies -Commentaries -Reviews
Language	-English -already translated into English	Studies in languages other than English Studies not translated in English
Time frame	open – November 2014	Post Novemebr 2014
Setting	High fidelity simulation lab	Healthcare agency site

SR objectives.

The objectives of the SR are stated such that they reflect the research question ensuring that the same elements (the population, intervention, comparative intervention, outcomes, and study design) (PICOTS) are present in the statements. Grimshaw (2010) asserts that the more specific the objectives, the more amenable the question is to a SR methodology. The objectives of the proposed SR are:

Primary objective:

To identify and assess the outcome measurement instruments used in high fidelity SBCE for the outcomes of competence, CT, CJ, and CR with undergraduate nursing students.

Objective 2:

To identify and map the most commonly used outcome measurement tools by type of SBCE scenario and by country of use (i.e. Canada, U.S.A., U.K., and Europe).

Objective 3:

To characterize outcome measurement tools by the type(s) of learning domain(s) depicted in the SBCE student outcomes.

Objective 4:

To characterize the outcome measurement tools by the outcomes measured in the SBCE.

Objective 5:

To characterize outcome measurement tools by the methods used to establish their reliability and validity. Following this, to determine which outcome measurement tools are most valid.

SR search strategy.

The purpose of the search strategy is to generate a comprehensive list of studies appropriate for answering the research question. The search strategy is a critical stage of the review as the validity of the review results is directly linked to the thoroughness of the search and its ability to locate all relevant studies (published and unpublished) (Bettany-Saltikov, 2010). The search also has value in that it locates current knowledge relevant to the concepts and contexts regarding what is known in a particular field (Petticrew and Roberts, 2006).

The search strategy outlines the intended approach to searching the literature and documents all sources to be searched. In particular, the strategy describes which search databases and sites will be used: bibliographic and subject databases, general search engines, theses and dissertations databases, grey literature databases, as well as journal databases. Non-bibliographic database sources are also described (hand searching journal issues and conference proceedings) as these unpublished articles may be robust enough to provide valuable information (Bettany-Saltikov, 2010). In contrast, Whittemore and Knafl (2005), caution that unpublished studies identified through conference abstract proceedings or through networking may not have the same methodological rigor as peer-reviewed publications and is not considered acceptable by all. Ancestry searching, journal hand-searching, and searching research registries are approaches also recommended for searching the literature (Whittemore & Knafl).

It is necessary to search widely because not all research is published in journals, and not all research published in journals is indexed in major databases, thus the research may not be easily retrievable (Higgins & Green, 2011). Another reason for searching widely is the long wait for publication after a conference presentation, (due to process of submission, peer review and required amendments). Thus, finding an abstract or poster presented at a conference will give some information, though limited (Bettany-Saltikov). This wide search aims to limit publication bias where studies with positive results tend to be published more frequently than those with negative results and where high prestige organizations tend to be published rather than those from lower prestige organizations (Webb & Roe, 2007).

The proposed SR search strategy includes four main categories of sources to provide a SR that is as comprehensive as possible and include: online databases, journal articles, grey literature and books (Bettany-Saltikov, 2010). This SR will search the following bibliographic databases: CINAHL (Cumulative Index to Nursing and Allied

Health), PubMed (Medline); EMBASE (Excerpta Medica Database) (strong in capturing conference abstracts, via personal communication with librarian Robin Parker February 27, 2014), ERIC (Education Resources Information Center), and ScienceDirect. In order to capture dissertations and theses, the ProQuest database will also be searched.

Subject specific database searches will include SIRC (Simulation Innovation Resource Center), and INACSL (The International Nursing Association for Clinical Simulation & Learning), and SSIH (the Society for Simulation in Healthcare). The Cochrane CENTRAL register and the Joanna Briggs Institute register will also be searched as searching research registries provides information about studies regardless of the statistical significance of findings because registration occurs before study completion (Conn et al., 2003).

Searches via general search engines have little empirical evidence to back up their value in providing potential studies (Higgins & Green, 2011). However, this SR will search Google Scholar, Intute and the Turning Research into Practice (TRIP), all of which are identified by the Cochrane Collaboration (Higgins & Green) as the best of these sources.

Journal articles are usually primary sources of studies and are the most up-to-date sources of peer reviewed articles and advances in the field of interest. Most of the database searches listed above will either retrieve the article or direct the researcher to other primary sources of literature (Bettany-Saltikov, 2010), thereby making a hand search of the relevant journals of questionable value. Higgins and Green (2011) also state that reviewers are not routinely expected to hand search journals for their reviews but they should discuss with their Trials Search Co-ordinator or supervisor whether it might be beneficial, bearing in mind that it may yield publications that have not yet been indexed in computerized databases. This is however, a labor and time-intensive task. (Conn et al., 2003). This SR will include hand searches of only the 2014 electronic issues of *Simulation in Healthcare* and *Clinical Simulation in Nursing* journals due to the time limitation of the study.

Ancestry searches of relevant articles will also be carried out. Ancestry searching involves reviewing the reference lists of relevant articles to locate other primary reviews thus expanding the number of eligible studies. It is important to note that relying solely on ancestry searching without adequate attention to other search strategies may yield a biased set of studies. This is because studies with statistically significant findings are more likely to be cited in reference lists than are studies without significant findings (Conn et al., 2003).

A grey literature search will be conducted to retrieve relevant conference proceedings, published abstracts, newsletters, and other unpublished material via the Canadian Electronic Library (Canadian Health Research Collection), Mednar.com, and by searching simulation newsletters and conference proceedings.

Books titles for nursing and simulation will be searched in the electronic databases for new releases (Bettany-Saltikov, 2010). Books are considered a secondary source of information for developing the background of the study but may yield simulation evaluation tool information to guide the reviewer to a primary source.

Search terms.

Search terms are also defined in the protocol keeping in mind that the search must be both sensitive and specific. Sensitivity in SRs refers to the comprehensiveness of the search to include all necessary and relevant studies, while specificity (precision) focuses on the retrieval of only relevant reports. The challenge in a SR is to achieve a balance between the two when developing the strategy, whereby the search terms are modified based on what has already been retrieved (Higgins & Green, (Eds.), 2011). At some point, each new search returns fewer relevant references and at this point, the researcher must determine whether it is plausible to continue the search (CRD, 2009).

Search terms for the review are focused on the intervention, outcome measure, and type of study in the review PICOTS. Higgins and Green (2011) assert that "it is usually unnecessary, and even undesirable, to search on every aspect of the review's (clinical) question" (p 6.4.2), as the population, setting, and outcomes may not be well described in the study abstract or title thus, are often not well indexed in the database. Each database will be searched using keywords and MeSH terms as appropriate. The search terms for the SBCE intervention are high-fidelity, simulation, simulat* clinical, clinical, and human patient simulators. Search terms for the type of study are randomized controlled trial, experimental, quasi-experimental, controlled trail, comparative, pilot study, and observational. The phrases teaching, simulation evaluation instruments, simulation outcomes measurement, and nurs*, evaluat*, and educat* will be used to narrow the search in order to capture studies relevant to nursing and those involving competence measurement instruments.

SR study selection protocol.

It is recommended, that more than one reviewer carry out the study quality assessment to help minimize bias and error, and allow for inter-rater reliability to be assessed (CRD, 2009; Webb & Roe, 2007). Two reviewers are a minimum expectation, functioning as a primary and secondary reviewer (Joanna Briggs Institute (JBI), 2014) but teams also carry out SRs (e.g. Cochrane Collaboration systematic reviews) (Higgins & Green, 2011). Alternately, Polkki et al. (2014) state that there is no agreement about the number of assessors for reviewing the methodological validity of a study.

In the case of two reviewers, the primary reviewer initiates the review; assigns the secondary reviewer to the review; is able to add, edit or delete their own review; determines the time frame of the review; critically appraises potentially relevant papers; provides an overall appraisal of papers following critical appraisal by the secondary reviewer, and conducts the primary data extraction from included papers (JBI, 2014). The secondary reviewer assesses every paper selected for critical appraisal, and assists the primary reviewer in conducting the review. Associate reviewers may be added to the review to extract data (with, in most cases, the secondary reviewer) from included papers. (JBI).

The researcher now designs a study selection form as part of the SR protocol and tests it to ensure its appropriateness. The study selection form identifies the PICOTS elements of the SR research question, the predetermined inclusion criteria, and study limits. This ensures that only studies meeting the inclusion criteria are selected because a single failed eligibility criterion is sufficient to exclude that study (Higgins & Green, 2011). Therefore, eligibility criteria are listed in order of importance on the form so that the first ‘no’ response can be used as the primary reason for exclusion of the study and the remaining criteria need not be assessed (Higgins & Green).

This SR includes two reviewers for the study selection and study quality appraisal phases. The study selection strategy and selection form will be piloted on a random sample of studies to refine and clarify the inclusion criteria and ensure that the criteria can be applied consistently by more than one reviewer (CRD, 2009). Once the form is revised as necessary, both reviewers will systematically search the literature for relevant studies guided by the use of the SR study selection form. See Appendix B for the study selection form created for this SR.

The first phase of study selection involves the research reviewers systematically sorting through database titles and abstracts of all the articles retrieved by the search, and excluding irrelevant studies. Irrelevant studies are filtered out by scrutinizing all the articles by title and abstract and selecting those that meet pre-determined criteria, leaving three categories of articles: a) those that will definitely be included, b) those that may or may not be included, and c) those that will be rejected (Bettany-Saltikov, 2010). Rejected citations fall into two main categories; those that are clearly not relevant and those that address the topic of interest but fail on one or more criteria such as population (CRD, 2009). Citations that fall in the second category are recorded with the reason for failure to meet the inclusion criteria as this increases the transparency of the selection process (CRD). When the article title or abstract is not sufficient to determine whether or not the study meets inclusion criteria the reviewers review the full text of the article to make the decision. The reviewers then closely examine the studies that were accepted and those that were unclear by reading the full text copies of each and then deciding if they meet the inclusion criteria (Bettany-Saltikov, 2010). The protocol specifies the process by which decisions on the selection of studies will be made including the number of researchers who will screen titles and abstracts, then full papers, and the method for resolving disagreements about study eligibility (CRD).

Both the primary and secondary reviewer for this SR will screen titles, abstracts and full papers for inclusion (according to the study selection form and protocol) and any disagreements regarding study inclusion will be discussed and, where possible, resolved by consensus (after referring to the protocol) (CRD, 2009). If necessary a third person will be consulted. Once this is completed the included studies are appraised for methodological quality.

SR study quality appraisal.

Phase two involves appraising the methodological quality of the studies assessing their internal validity (degree to which the study design, conduct, analysis, and presentation minimize bias) and external validity (the generalizability of the study findings). The appraisal determines the degree to which the study is free from methodological errors, the occurrence of which would reduce the researcher's confidence about the studies' potential validity (Grimshaw, 2010; Petticrew & Roberts, 2006). Quality assessment of studies is also referred to as the assessment of risk of bias and ultimately helps the reviewer decide if the studies are robust enough to guide stakeholder decisions (CRD, 2009; Moher et al., 2009). Although quality assessment can be used to exclude studies that do not meet certain criteria, this is not standard practise (CRD, 2009). Not removing the study at the quality appraisal stage allows the primary reviewer to consider threats to validity during the analysis and interpretation phases of the SR thus allowing the researcher to explain or interpret differences across studies and inform a qualitative interpretation of the risk of bias (CRD; Grimshaw),

It is at this stage of the quality appraisal protocol that the researcher plans how the studies will be assessed for internal and external validity, identifies the scale or checklist for quality assessment, and again determines how disagreements between reviewers will be handled (Bettany-Saltikov, 2010).

This SR will include parallel independent quality assessments of studies by the two reviewers to minimize the risk of errors. Once again, any disagreements between assessors regarding study quality and inclusion will be resolved according to the predefined strategy of reaching consensus (CRD, 2009).

The researcher keeps in mind four critical stages in experimental studies where bias or error occurs: a) the allocation of participants to study groups where selection bias may occur but can be minimized by randomization; b) any differences in care provided to participants in the study groups, other than the intervention being evaluated resulting in performance bias, (which can be minimized by blinding participants and care providers to the treatment group); c) any differences between groups in the number of participants who did not complete the study (attrition bias); and d) differences in the assessment of outcomes of participants in the different study groups (minimized by blinding the assessors to study group allocation) and is detection bias (Evans in Webb & Roe, 2007).

Study quality appraisal also includes the relevance of the research question, the appropriateness of the data analysis and presentation, and any ethical implications. This appraisal phase is critical to the evidence and quality of the SR as both depend directly on the quality of the studies included (Bettany-Saltikov, 2010). As well, determining the differences found in the study's quality may explain the differences in the study's results thus providing guidance to the researcher in interpreting the results for the purpose of the review question.

Many methods exist to assess study quality and no single approach is appropriate for all SRs (CRD, 2009), thus an important decision at this protocol stage is to determine which appraisal scale or checklist is appropriate for the SR. A search and review of ten nursing SR articles between the years 2004-2014 showed that five studies did not mention use of any quality appraisal tool and each of the other five used different quality appraisal tools: the Critical Appraisal Skills Programme tool (Cant & Cooper, 2009), a JBI tool (Lapkin et al., 2004) , the Jadad tool (Yuan, Williams, & Fang, 2011), the McMaster quantitative tool (Liu, Cheon, & Thomas, 2014), and a researcher developed qualitative tool (Shearer, 2013).

The choice of an appraisal tool is guided by the designs of the primary studies, the level of detail required in the appraisal, and the ability to discriminate between risk of bias (internal validity) and generalizability (external validity) (CRD). The appraisal scale or checklist becomes the record of the study characteristics, quality and results. Data extraction is linked to assessment of study quality in that both processes are often undertaken at the same time (CRD).

The quality indicators used to appraise articles in this SR include the quality of the study information on: choice of the outcome measurement tool, outcome measurement tool validity and reliability, outcome methods, participant eligibility and recruitment, setting and locations of study, intervention method, sample size, reported statistical methods of analysis, reported outcomes with measurements, interpretation of results, and study generalizability (CRD, 2009).

The quality assessment tool which best fits these indicators and is the method selected for this SR is the COnsensus-based Standards for the selection of health

Measurement INstruments (COSMIN) (Mokkink et al., 2010). The COSMIN initiative aims to improve the selection of health measurement instruments by the use of the COSMIN critical appraisal tool (checklist) containing standards for evaluating the methodological quality of studies on the measurement properties of health measurement instruments.

The COSMIN checklist was developed by an international Delphi study via a multidisciplinary, international collaboration including involvement of 43 experts. Consensus was reached on standards for design requirements and appropriate statistical methods for assessing measurement properties. The COSMIN checklist was developed based on these standards. The focus was on health-related patient-reported outcomes but the authors assert that the checklist is also useful for evaluating studies on performance-based instruments or clinical rating scales (Mokkink et al., 2010). It is particularly useful in SRs to appraise the methodological quality of studies on measurement properties, thus making it appropriate for this SR on outcome measurement instruments. The COSMIN steering group asserts, "instrument selection should be based on systematic reviews in which the content and measurement properties (reliability, validity, responsiveness) of all instruments measuring a certain construct, are critically appraised and compared" (COSMIN checklist/ background). In an SR of measurement properties of primary studies, the COSMIN steering group considers the rating of the methodological quality of these studies to be an important step.

The COSMIN checklist is comprised of 12 boxes used to assess whether a study meets the standards for good methodological quality. Two boxes evaluate whether general requirements of a study on measurement properties are met. Nine boxes evaluate the quality of the assessment of different measurement properties concerning: internal consistency, reliability, measurement error, content validity (including face validity), construct validity (subdivided into three boxes, structural validity, hypotheses testing, and cross-cultural validity), criterion validity, and responsiveness. Finally, one box is used to evaluate the quality of a study on interpretability of a measurement instrument. While interpretability is not considered a measurement property it is an important requirement for the suitability of an instrument in research or clinical practice (Mokkink et al., 2010).

The researcher employing the COSMIN checklist is aided by a detailed user's manual, a taxonomy, and an optional checklist with a four point rating system which gives each study an overall quality rating. The taxonomy includes all measurement properties that should be evaluated when an instrument is used for evaluative purposes (Mokkink et al., 2010). The COSMIN items in this tool have four response options in order to increase the discriminative ability of the items and to represent excellent, good, fair, or poor methodological quality. The methodological quality score per box is obtained by taking the lowest rating of any item in a box (Mokkink et al., 2010). The COSMIN developers recommend using this scoring system when performing a systematic review of measurement properties. Alternatively, Grimshaw (2010), Higgins and Green (2011) and the CRD (2009) discourage creation of an overall quality score for each study. They note that many scales have been poorly validated and include items not directly related to internal validity (Grimshaw; Higgins & Green) and that the weighting assigned to various items varies between scales and does not usually account for the direction of the bias (CRD).

The COSMIN checklist is recommended for researchers conducting a SR on measurement properties by the COSMIN steering group, who have expertise on measurement properties of health related outcome instruments. Additionally, the four-point checklist has specific criteria for each rating choice providing clarification for which rating level is most appropriate (Mokkink et al., 2010) decreasing the subjectivity of the rating by the reviewer.

Additionally, the COSMIN checklist is the only quality appraisal tool with the design: to evaluate primary studies on performance-based instrument measurement properties as well as study quality; that is not constricted to one type of study design (important because this SR is likely to include various study designs); that has been used extensively with SRs on measurement properties; and that has clearly defined criteria for each property. Furthermore, the COSMIN purpose and definition of a SR of measurement properties (an SR in which the content and measurement properties of measurement instruments are critically appraised and compared) match the objectives of the proposed SR (Mokkink et al., 2010).

SR audit trail.

The study selection process is documented, detailing reasons for exclusion of studies that are ‘near-misses’ (CRD, 2009). Houser (2012) suggests a process for reviewers making decisions about the inclusion of studies in the final recommendation stage: 1) apply search criteria and develop a master list of citations which results in a list of potential relevant citations after screening; 2) review abstracts and citations that were excluded with reasons which results in a list of studies retrieved for more detailed evaluation; 3) create a list of studies excluded after full review of article with reasons;

and 4) result is a final list of citations for inclusion in final review. Houser states that it is not unusual for a large number of potential citations to be reduced by more than half due to the inflexible inclusion criteria and rigorous methodological quality that is required of studies in the review.

In the proposed SR a record will be maintained of each abstract and article reviewed, as well as justification for any study eliminated to provide transparency of key decisions thus ensuring that the reviewer is using objective and defensible reasons for selecting studies for final inclusion as well as minimizing the potential effects of researcher bias on the study outcome. A flow chart adapted from the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) will depict the flow of SR study inclusion steps from article identification with sources and number retrieved, screening numbers and rationale, appraisal eligibility numbers and rationale, and final number of included studies (Moher, Liberati, Tetzlaff, & Altman, 2009). See Diagram 1.

SR data extraction.

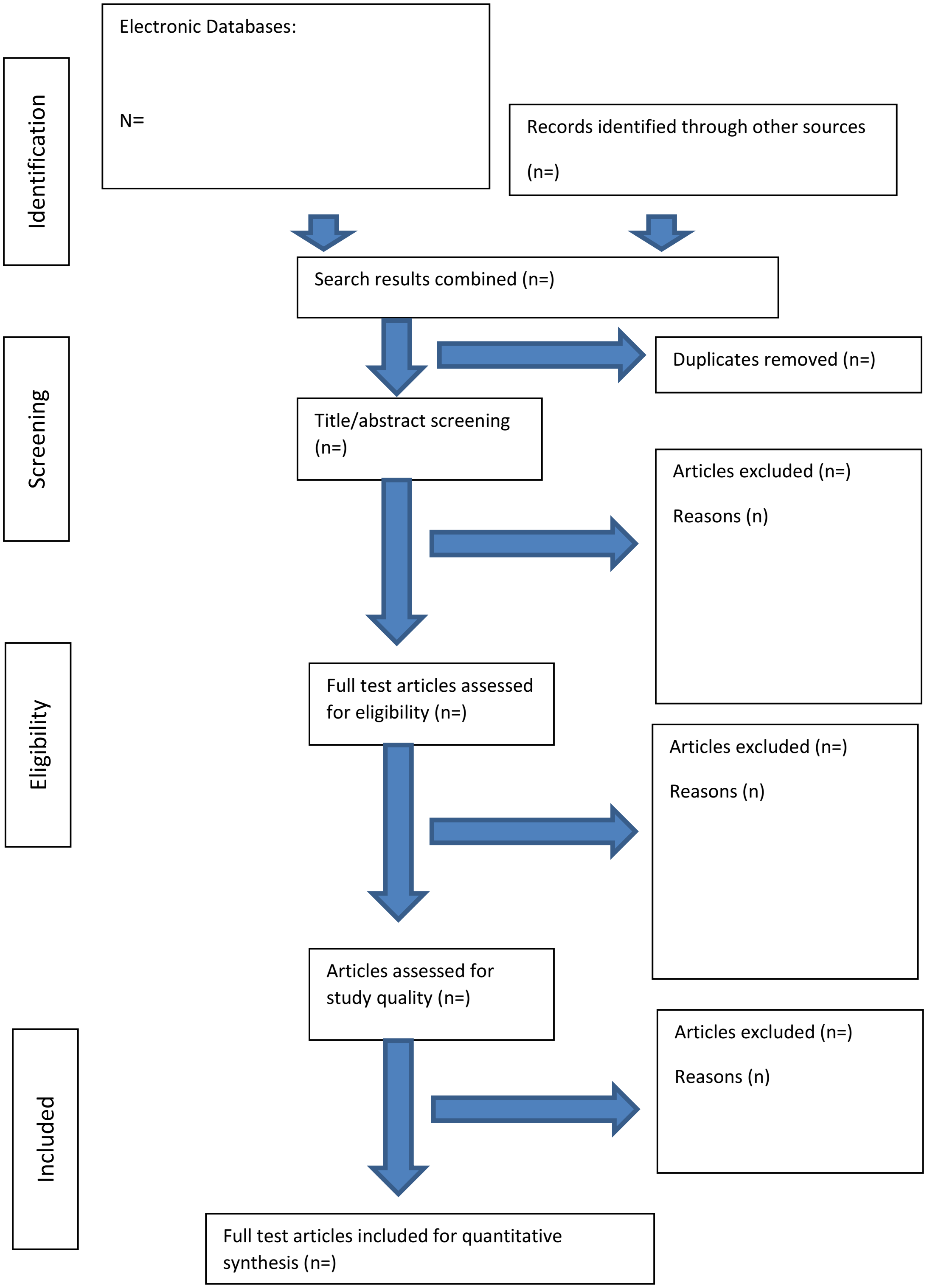
Phase three involves extracting data from the studies by searching each study for the PICOTS elements. The first step is to plan the type of analyses and list the tables that will be included in the report as this helps to identify which data should be extracted. (CRD, 2009). The process of data extraction is standardized to improve the validity of the review results by comparing each study to a data extraction form that was developed and piloted on one or two of the articles by the reviewer. This assures that the form is useful and appropriate for capturing the required details (Bettany-Saltikov, 2010). Standardized data extraction forms can provide consistency in a systematic review, while at the same time, reducing bias and improving validity and reliability (CRD). As well, the data extraction form serves as a direct link to the SR review question, as a record of this phase of the review, and is the data that will be analyzed (Evans in Webb & Roe, 2007). The data extraction stage involves going through each of the primary articles included in the final list and highlighting all relevant information that will answer the research question, then extracting this information to the data extraction form (Bettany-Saltikov, 2010). See Appendix E for the data extraction form for this SR.

Data analysis and synthesis in the SR protocol.

The next stage of the protocol is determining the data analysis method most appropriate to answer the type of review question and the type of primary studies to be included in the review (CRD, 2009). A narrowly focused review question may permit statistical analysis, while a broader question may necessitate the use of tables and narrative summaries (Bettany-Saltikov, 2010). There is agreement that a SR can, and should, use a range of different research designs, yet there remains a lack of clarity on how best to synthesize the results. The results of the studies may be presented in a SR as conclusions in a simple summary of results; by a meta-analysis of quantitative results; or by a narrative synthesis of results where the reviewer identifies certain elements of interest, then compares and combines results, thus integrating them to provide a new interpretation (Bettany-Saltikov, 2010).

Meta-analysis is most commonly used with systematic reviews that typically include only RCT studies and even then, it is necessary that all the studies have the same question, administer the intervention in a similar manner, have a similar population, measure the same outcomes for all participants, and use the same research design (Evans in Webb & Roe, 2007).

Diagram 2. PRISMA Flow Chart



Flow diagram showing article selection according to PRISMA (preferred reporting items for systematic reviews and meta-analyses) criteria (Moher, Liberati, Tetzlaff, & Altman, 2009).

When primary studies cannot be compared statistically, a narrative analysis of quantitative data is acceptable (CRD, 2009); there are SR topics where it can be decided *a priori,* that a narrative approach is most appropriate (CRD). The *a priori* approach is appropriate when the researcher anticipates that the topic under review will likely result in diversity in the included studies in terms of settings, interventions, and outcome measures (CRD). The PRISMA guideline for reporting SRs notes that the decision of whether or not to combine data statistically involves statistical, clinical and methodological considerations (Moher et al., 2009). The clinical and methodological decisions lie with the reviewer (and team) and may be more subjective than the statistical decision, which is "more technical and evidence-based" (p.33) (Moher et al.).

The proposed SR will utilise a narrative synthesis rather than a meta-analysis of the SR data for several reasons. Firstly, it is anticipated that the primary studies will be of varying design (with very few experimental designs), thus making pooling of the data inappropriate. Also, variation in the study outcomes and in the type of outcome measurement instruments is anticipated which also makes statistical analysis inappropriate.

Narrative synthesis protocol for proposed SR.

Synthesis is defined as the collation, combination, and summary of the findings of individual studies included in the SR where the defining characteristic is "the adoption of a textual approach that provides an analysis of the relationships within and between studies and an overall assessment of the robustness of the evidence" (p.48) (CRD, 2009). Grimshaw (2010) notes that despite the fact that narrative synthesis is one of the most common approaches to synthesis there remains surprisingly little guidance to their conduct. SR literature indicates that narrative synthesis has typically not followed a strict set of rules, yet bias must be minimized as with every other stage of the SR process (CRD, 2009). Narrative synthesis is inherently a more subjective process than meta-analysis, thus the researcher's approach must be rigorous and transparent to reduce the potential for bias. This enables reliable conclusions to be drawn from the assembled body of evidence (CRD).

General frameworks and guidelines exist to help the reviewer maintain transparency and add credibility to the SR process (CRD, 2009). From the work of Petticrew and Roberts (2005), Grimshaw (2010) reports three stages of the narrative review as 1) organizing studies into logical categories (e.g. study design, outcomes) to guide analysis, 2) providing a narrative description of the findings of each study with the description of study quality (termed within-study analysis), and 3) providing an overall summary of the study findings considering variations in study quality and other variations (e.g. variations in population, intervention, and settings) that may affect generalizability of studies (termed cross-study synthesis).

A general framework that presents four elements of a narrative synthesis includes: the development of a theory of how, why, and for whom the intervention works; the development of a preliminary synthesis of findings of included studies; investigating relationships within and between studies; and evaluating the rigour of the synthesis (CRD, 2009). In this model the researcher moves iteratively between the four elements choosing tools and techniques that are appropriate to the data being synthesized and providing justification for these choices. Not all four elements are necessary in every SR. No matter which method is followed, the descriptive process is to be both explicit and rigorous and decisions about grouping and tabulating data is based on the review question and what has been planned in the protocol (CRD).

Proposed SR synthesis protocol.

The narrative synthesis for the proposed SR will follow the three stages outlined by Grimshaw (2010) and will begin by organizing the studies into logical categories by: study design, outcome measurement instrument type, domain of learning, and type of SBCE clinical scenario. The data extraction form and the COSMIN quality appraisal form will be the source of this information from which the reviewer will identify the data for each category. Once this is complete, a table portraying a descriptive summary of the included studies will be developed. The table will include the descriptive characteristics of all included studies (authors, title, study design, intervention method, participants, measurement instrument, outcomes, country of instrument use, and results).

This addresses the primary objective of identifying the outcome measurement instruments used in high fidelity SBCE for the outcomes of competence, CT, CJ, and CR with undergraduate nursing students. Additionally, it addresses part of objective two as the table will identify the country of use of the measurement instrument. Finally, it enables the SR to meet the expectation that all systematic reviews should begin with text and tables providing an initial descriptive summary and explanation of the characteristics and findings of the included studies (CRD, 2009). As the proposed SR doesn't involve re-calculating summary statistics, but rather relies on the reported results of the author's analyses, these results/findings will be included in this table (CRD).

Data synthesis for objective three (to characterize outcome measurement tools by the type(s) of learning domain(s) depicted in the SBCE student outcomes) will be conducted by identifying the learning domain associated with the measurement instrument and outcomes from each study. This data will be drawn from the study data extraction sheets and then summarized in a descriptive table including the study name, instrument type (e.g. rubric, checklist), instrument name, and domain of learning. This same table will include a column characterizing the outcome measurement tools by the outcomes measured in the SBCE (competence, CT, CR, and CJ), thus incorporating the findings for objective four.

Objective five concerns characterizing outcome measurement tools on the methods used to establish their reliability and validity, then to determine which outcome measurement tools are most valid. Data regarding the validity and reliability of the reviewed measurement instruments will be extracted from the papers during data extraction and recorded therein. The validity and reliability (as reported by the study authors) of each instrument will then be synthesized in a table and discussed in the results of the proposal.

In order to determine whether the reported reliability and validity of the measurement instruments in the primary studies meet the standards of rigour for these properties, the researcher will compare the reported reliability and validity of each study with the expected reliability and validity requirements for measurement instruments. The expected reliability and validity requirements for measurement instruments that will guide the SR comparison and discussion are as follows.

Reliability of measurement instruments

Reliability affects the precision of a measure such that when measures are reliable the instrument is consistently measuring the characteristic of interest (Shelestak &

Voshall, 2014). Validity enhances the accuracy of a measure such that when an instrument is valid it is known to represent the underlying attribute well (Houser, 2012). Validity increases the credibility of the conclusions and supports application of the results to practice. It is possible that an instrument can have reliability without validity so studies may report instruments with documented reliability without reported validity (Polit & Beck, 2008). Instruments cannot be valid without being reliable, therefore instruments reporting validity but not reliability will be highlighted in the discussion of results as both are required to trust the outcome of a study (Houser, 2012; Polit & Beck in Frasure, 2008).

Reliability analysis and the statistics associated with it ensure that the instrument is stable and verify the degree to which an instrument is stable internally among participants, between raters, and over time. The three types of reliability are: stability, internal consistency, and equivalence (Polit & Beck, 2009).

Stability reliability.

Stability is the extent to which scores are replicated on separate testing situations (Polit & Beck, 2008). Test-retest reliability determines the stability of the instrument over time indicating such that when the instrument is reliable and is used repeatedly, the results are due to actual changes in the subject, not the instrument. Stability is measured by a test-retest correlation coefficient which should exceed or equal 0.7 (some argue 0.5) for this measure (Houser, 2012; Shelestak & Voshall, 2014).

The prosed SR will capture each included instrument's test-retest reliability by searching the primary study for the correlation coefficient as identified by the study authors. The reported correlation coefficient will be assessed against the recommended correlation coefficient standard to determine if the study coefficient meets this standard. The study's correlation coefficient will also be displayed in a table that will list instrument validity and reliability. The interpretation and significance of the studies' reported stability correlation coefficient will be discussed in the individual summary of each study.

Internal consistency reliability.

Internal consistency is the degree to which the individual instrument items measure the construct of interest. The researcher will identify and report the studies' internal consistency as measured by and reported as the alpha coefficient statistic, Cronbach's alpha. This measure represents the extent to which the variability on individual tool items represents the variability in the overall instrument (Houser, 2012). It is most commonly reported for the tool subscales (some studies may have Cronbach's alpha for the overall tool as well) and is the most widely used method for evaluating internal consistency reliability, where highest values represent the strongest internal consistency reliability (+1.00 – 0.00 lowest) (Frasure, 2008; Polit & Beck, 2008). Ideally, the coefficient alpha should meet or exceed 0.7 as a minimum. Moderate reliability measures are 0.7 to 0.9 and 0.4 is considered unacceptable (Houser). In the event of high-stakes testing the correlation coefficient should be 0.9 or greater (Shelestak & Voshall, 2014). A strong internal consistency and/or test-retest consistency is important when assessing student performance during simulation (Shelestak & Voshall).

The proposed SR will capture and report the outcome instrument's internal consistency reliability for each included primary study. The alpha coefficient statistic, Cronbach's alpha, will be identified as reported in the primary study and then reported in the instrument validity and reliability table. The SR reviewer will assess whether the reported alpha coefficient statistic meets instrument minimum, moderate, or maximum or if it, in fact, falls below recommended standard of Cronbach's alpha 0.7. If no internal consistency reliability is reported in the study, this too will be noted as absent in the instrument table, as will any reported measures. The reviewer will note whether the coefficient is reported for the tool subscales as well as the tool itself. The SR individual study summary will present the discussion of the tool's alpha score and it's meaning to the overall reliability of the instrument. This is especially important as a strong internal consistency is critical for instruments measuring student performance during a SBCE to ensure that each time the tool is applied it is correctly measuring the same construct for each student.

Equivalence reliability.

Equivalence reliability, known as inter-rater reliability (IRR), quantifies the stability of a measure across raters (whether two or more raters agree on the ratings). IRR is an essential component of high-stakes evaluation and should be assessed before using the tool in a testing situation (Shelestak & Voshall, 2014). The degree of agreement should be documented in the primary study by simple percentage agreement (of 0.85 or greater) or by Cohen's kappa (considered preferable). The kappa value should be equal to or greater than 0.85 with a p value less than 0.05, indicating that the agreement was not by chance (Houser, 2012; Shelestak & Voshall). Cohen's kappa is a more rigorous measure of IRR with dichotomous data (yes/no; pass/fail) and is essential for summative and high-stakes testing (Shelestak & Voshall). For ordinal, interval, or ratio data in instrument scores the most common statistic used is the intraclass correlation (ICC), where a rating of one indicates perfect agreement and zero indicates that agreement occurred by chance (Shelestak & Voshall). Values of 0.75 – 1.0 represent excellent agreement, and 0.04 or less is considered poor.

The proposed SR will capture the equivalence reliability, or inter-rater reliability (IRR), by examining each primary study and extracting the reported IRR for each study onto the SR data extraction sheet. The reviewer will then report the IRR by the percent agreement or Cohen's kappa, as reported by each studies' researcher(s). The reported IRR will be entered into the instrument reliability and validity table; then assessed and compared narratively with the recommended IRR standard in the summary of individual studies. The same process of analyses will be carried out for the ICC statistic.

Studies should report at least one test of reliability for the instrument used. The gold standard is to assess consistency within the instrument and among individuals over time, but this testing is dependent on study resources. Instruments developed by the researcher should conduct and report a pilot test on a group of subjects to assess reliability (Houser, 2012).

The proposed SR will assess and discuss the reported number and type of reliability tests carried out for each instrument, and reported testing of the tool over time, or failure to do so with reasons such as reported study resource limitations. The reviewer will also identify from each study if any pilot testing was done with the instrument before applying it to a new population, setting or construct. Both of these standards will be discussed in the individual study summary as well as implications for the tool in the overall summary of the findings.

Validity of measurement instruments.

Content validity.

Validity is the ability of an instrument to consistently measure what it is supposed to measure and is harder to test than reliability as it must be tested and retested to ensure it is effective across settings and situations (Houser, 2012). Ideally, each study will report two to three validity measures: face, content, criterion-related, and construct. Content validity involves a subjective judgment about whether an instrument measures the designated construct (Shelestak & Voshall, 2014). Reported content validity may mean that face validity has been assessed (i.e. it looks like the tool measures xyz) or that a panel of experts (i.e. those with a strong familiarity of the concept) has verified that it measures the correct concepts. Content validity may also be established through a review of the literature on the concept, and from the findings of a qualitative study providing representativeness of the population on the experience (Houser, 2012). The most common best result is that a panel of experts evaluated the fit of the tool with the underlying concept(s) and possibly with a test blueprint. The primary study should report that the instrument was subsequently updated according to the recommendations of the experts. A more precise measure of content validity is the content validity index (CVI) calculated during the development of the instrument (Shelestak & Voshall). It too, is based on multiple experts' rating of relevance of the instrument where each expert is asked to numerically rate the relevance of each instrument item from not relevant (0) to very relevant (3). The level of agreement for each item is then calculated.

The reviewer of the proposed SR will gather all content validity data from each primary study as reported by the study author(s). The reviewer will note which method(s) the authors used to ensure content validity (i.e. expert panel review, literature review, CVI, qualitative findings) and report this in the instrument validity and reliability table. The analysis of the reported content validity method(s) will be discussed in the individual study summary comparing study results to recommended standards and methods for content validity in instruments. The implications of the content validity findings for each instrument will also be discussed in the study summary.

Criterion-related validity.

Criterion-related validity involves determining the relationship between the instrument and an external criterion (Houser, 2012; Polit & Beck, 2008). The instrument is considered valid if its scores correlate highly with the scores on the criterion. A requirement for this measure is a criterion that is available, reliable, and valid to serve as a comparator. Once the criterion is selected a correlation coefficient is computed between the scores on the instrument and the criterion, with a high coefficient (> 0.5) indicating strength of the validity (Houser; Polit & Beck). Criterion validity can be categorized into three types: concurrent, predictive, and discriminative.

Concurrent validity is measured when the instrument and criterion scores are collected at the same time, whereas predictive validity is established by applying correlation or regression statistical analysis to determine if the instrument is correlated to or can predict the construct of interest. Predictive validity is useful for instruments that predict future performance (e.g. new graduate competency measured against actual competency of a nurse) (Houser). Discriminate validity demonstrates the tool’s ability to discriminate between those who have a characteristic from those who don’t, and is considered valid if it accurately sorts subjects into classifications (e.g. those with a disease, those without).

The criterion-related validity data will be identified and extracted from each primary study and recorded on the SR data extraction form. The resulting data will be presented in the instrument validity and reliability table, then discussed in the individual study summary. The reviewer will note whether the study authors were able to identify and use a gold standard criterion for their construct of interest and, if not, whether the chosen criterion was considered a reliable and valid comparator. The reviewer will then determine whether the study authors chose concurrent, predictive or discriminate validity and whether this was appropriate for the type of study under review. Lastly, the reported correlation coefficient will be assessed to determine if it meets the standard requirement of >0.5. The overall significance of the results will be discussed in the study summary.

Construct validity.

The most important of these validity properties is construct validity as it indicates that a measurement captures the hypothetical basis for the variable, which is abstract and difficult to evaluate but very valuable (Houser, 2012). It ensures that the results will represent reality, which is fundamental to a strong evidence base as instruments cannot truly measure intended concepts without having evidence of construct validity (Frasure, 2008.) Factor analysis for subscale structure is a common method of construct validation that groups items within the instrument according to shared variability.

Construct validity is important in the studies included in this SR because it is best suited to instances when test scores assess attributes that are not easily or objectively measured (i.e. competence). As well, the researcher will examine the study to see if the authors report if and how the instrument has been reevaluated, if it is being used in a new problem, setting, or with a new population other than its intended use. Repeated use and testing also strengthens the tool's ability to apply to different problems. However, in dramatically different situations the tool should be tested with the new population and/or setting or before applying it (Houser).

Construct validity findings will be identified by the reviewer when extracting data from the study. The reviewer will look for factor analysis results, reevaluation of the instrument (if used in a new problem, setting or population), and reports of repeated use or testing of the tool. The factor groups will be analyzed to see if they represent the conceptual basis of the instrument and any repeated testing or re-evaluation of the tool will be noted. The reviewer will assess the method of instrument re-evaluation (in the instance of using the tool in a situation other than its intended use), and the appropriateness of the method and new situation. This analysis is done by evaluating the instrument's conceptual and operational definitions for fit with the new study. The results of the reviewer's analysis will be presented in the discussion summary of each instrument.

Stage 2 of SR Narrative Synthesis

The second stage of a narrative synthesis is a narrative description of the findings of each study as well as the description of study quality (termed within-study analysis). The researcher will follow Houser's (2012) validity and reliability recommendations in summarizing findings about the instruments by discussing whether: the instrument is clearly linked to concepts in the study, the instrument and measures are described objectively, the reliability of the instrument is described with supporting statistics, the validity of the instrument is described with supporting statistics, and whether a detailed protocol for the use of the instrument is provided.

Additionally, results from the COSMIN assessment tool regarding the validity and reliability of each tool will be summarized and presented along with the studies' methodological quality appraisal from the same COSMIN tool. The results of the study quality or risk of bias will also be presented as a separate table. The results include: choice of the outcome measurement tool, outcome measurement tool validity and reliability, outcome methods, participant eligibility and recruitment, setting and locations of study, intervention method, sample size, reported statistical methods of analysis, reported outcomes with measurements, interpretation of results, and study generalizability. The assessments from each COSMIN quality appraisal criterion will be reported, as reporting a score total lacks sufficient detail to describe where sources of bias may arise (Tricco, Tetslaff, & Moher, 2011).

Finally, according to stage three of the framework, the proposed SR will provide an overall summary of the study findings considering variations in study quality and other variations (e.g. variations in population, intervention, and settings) that may affect generalizability of studies (termed cross-study synthesis). This is in accordance with the expectation that the SR reviewer should discuss the quality, strength, and applicability of the evidence for each main outcome when summarizing the results (Tricco, Tetslaff, & Moher, 2011). The relevance of the results will also be considered for key stakeholders (e.g. educators, researchers, patients, health care providers) because this will help increase the applicability of the results for these groups.

Presentation of findings in SR protocol

The comprehensive search results may be presented in a table or narratively, with the goal to make the search transparent and replicable. This includes: databases and time frame searched, dates of search, number of hits, number of articles discarded, and number of articles retrieved. As well, the researcher decides how to present the details of each study in a table with a full discussion of each in narrative. The discussion includes the findings in relation to relevant background literature and the study aims, as well as presenting a comparison of findings and overall conclusions of the studies.

It is the intention that the reviewer adheres to the predetermined protocol, but this is not always possible or appropriate thus requiring changes. Higgins and Green (2011) state that it is important that protocol changes should not be made on the basis of how they affect the outcome of the research study. They stress that *post hoc* decisions made when the impact on the results of the research is known (such as excluding selected studies from a SR) are highly susceptible to bias and should be avoided.

Results to be presented include: results of the search, a summary of all included studies, a summary of the critiques of the included studies, and a summary of data extracted. This will be presented in narrative form with tables to complement the text and enhance its meaning. The condensed results of the studies will be presented in the characteristics of study table to ensure the validity of this SR (Whittemore & Knafl, 2005). This table will include: the authors, study title, methods, participants, measurement instrument, the method of observation, outcomes, results, and descriptive statistics of the study. Whittemore and Knafl also recommend strengthening the validity of the SR by presenting the credibility of the results of the primary studies as a way to evaluate the strength of their evidence. The results of the quality appraisal of each study will be presented in the table of study quality indicators which will include the study and date, design, sample size, presence of focused research question (y/n), selection

/allocation to groups (method), missing data, and comprehensiveness and clarity of reporting.

The instrument characteristics will also be presented in a table as well as discussed narratively**.** This table will include the author, instrument name, subjects, instrument description, scoring, reliability (Cronbach's alpha for internal consistency reliability and the p value for overall and subscales; test-retest reliability); and validity (as described in study narrative description). The competence, CT, CJ, and CR outcomes evaluated by the measurement instrument will be discussed narratively and presented in a table including the: study; clinical theme (e.g. medical-surgical skills, CT, CJ); and assessment instruments.

CHAPTER 4 FINDINGS

Introduction

The aim of the proposed systematic review (SR) was to determine which observational outcome measurement (evaluation) tools provide measures that are both valid and reliable in measuring undergraduate nursing students' outcomes of nursing following a high fidelity simulation-based clinical experience (SBCE).The outcomes under study are: 1) competence, 2) clinical judgment (CJ), 3) clinical reasoning (CR), and 4) critical thinking (CT). The SR research question was: What observational outcome measurement tools are reliable and valid in measuring students' individual competency outcomes for simulation-based clinical experiences using high fidelity patient simulators for undergraduate nursing student clinical experience? A comprehensive systematic search of the nursing literature framed by the question PICOTS elements guided the development of the SR study eligibility criteria and subsequent search strategy (CRD, 2009; Grimshaw, 2010).

The scope of this search was defined by the following study criteria: 1) English, (or translated into English); 2) search dates: from no specified start date to November 4 2014; 3) population: undergraduate nursing students in any year of their program in either an associate degree or baccalaureate degree-nursing program; and 4) an intervention that was identified as a high-fidelity SBCE inclusive of a high fidelity patient simulator and clinical scenario with a realistic environment and equipment. The geographical location of the studies was not limited but the setting was limited to a high fidelity clinical simulation laboratory or learning space. For a comparator intervention to be included in the study it had to include: a) an on-site clinical practicum (at a healthcare agency or community site), b) a lab-based clinical experience, c) a case study, d) simulated mannequin-based clinical scenarios, e) simulated standardized patient-based clinical scenarios, f) clinical scenario lecture classes, g) a low or medium fidelity SBCE, or h) low or medium clinical skill simulation experiences.

The criteria for study outcomes were identified as CT, CJ, CR, and/or competence measured by outcome measurement tools based on evaluator observation identifying these outcomes. The study design inclusion criteria comprised: primary research studies of experimental design (RCTs), quasi-experimental design (non-randomized controlled trial, comparative design before-and-after study, and interrupted time series), pilot studies of experimental or quasi-experimental designs, and observational pre and post cohort studies.

The analysis and synthesis of findings for this SR study follows the three stages of the narrative review as outlined by Grimshaw (2010) which include: 1) organizing studies into logical categories (e.g. study design, outcomes) to guide analysis, 2) providing a narrative description of the findings of each study with the description of study quality (termed within-study analysis), and 3) providing an overall summary of the study findings considering variations in study quality and other variations (e.g. variations in population, intervention, and settings) that may affect generalizability of studies (termed cross-study synthesis). Chapter 4 addresses stage one and two of the analysis of findings while Chapter 5 (the discussion of the findings) is stage three of this narrative model.

Structure of Study SR

This SR followed the recommended steps and structure of a SR to ensure a systematic and transparent methodology with objective findings (Bettany-Saltikov, 2010;

CRD, 2009; Grimshaw, 2010; Higgins & Green, 2011). The steps of the structure for this SR are presented in Appendix F, Table 3.

Search Terms

The major search terms used in the database searches were: outcomes education, outcomes assessment, teaching methods clinical, observational methods, student performance appraisal, reliability, validity, instrument creat*, measurement issues and assessments, critical thinking, clinical judg*, competency assessment, clinical N2 reason*, assess* instrument, and psychomotor performance. Appendix F, Table 11 presents a complete comparative list of search terms for CINAHL, Medline, Embase, PsycINFO and ERIC. In accordance with PRISMA guidelines for reporting SR results a full electronic search strategy for the CINAHL database appears in Appendix F, Table 7.

Systematic Review Study Selection Results

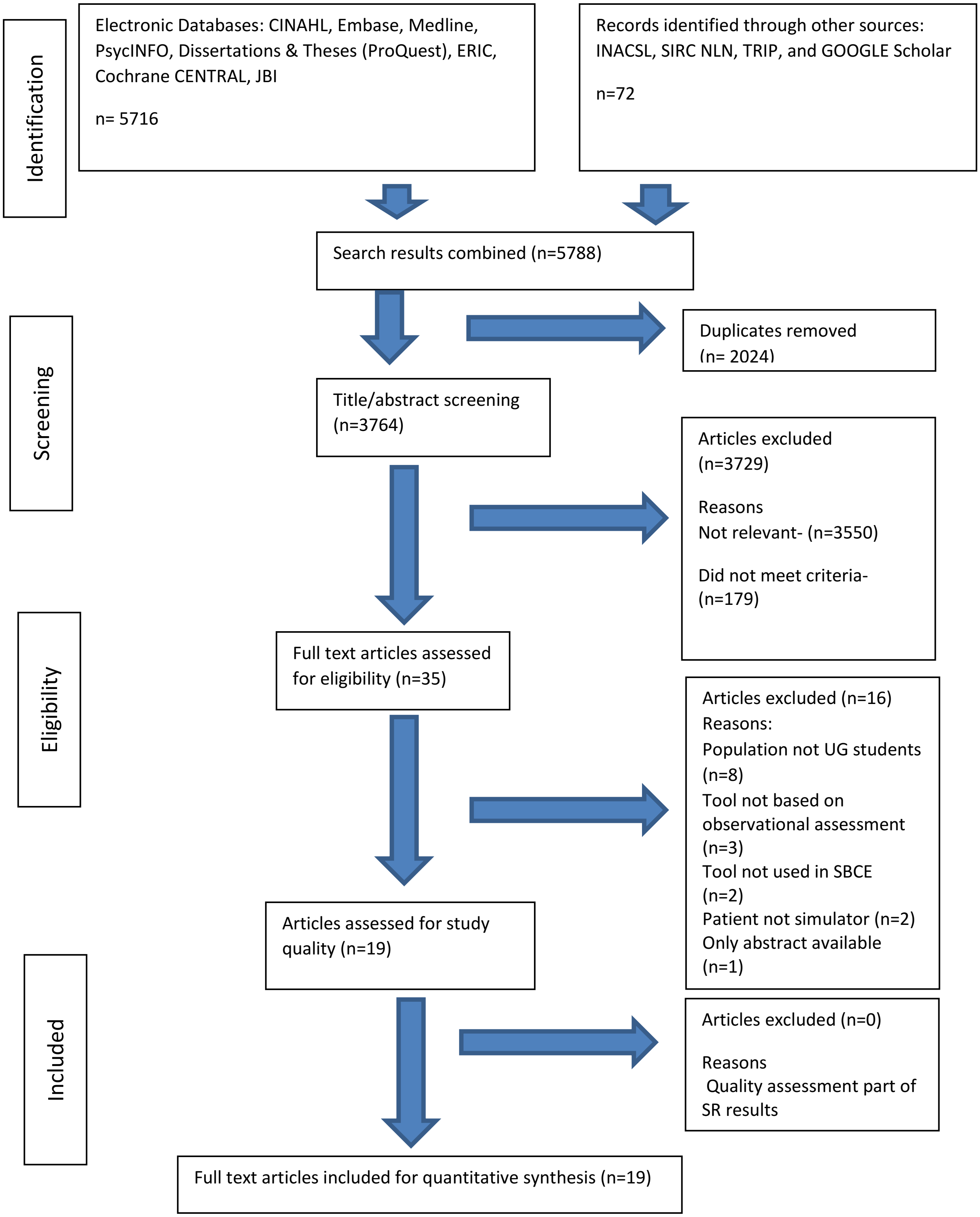
Eight databases and four search sites were searched systematically for studies related to SBCE and observational outcome measurement (evaluation) instruments. The databases included were: Cumulative Index to Nursing and Allied Health Literature (CINAHL), Excerpta Medica dataBASE (Embase), Education Resources Information Center (ERIC), Medline/ PubMed, PsycINFO, dissertations and thesis database (ProQuest), Cochrane Central Register of Controlled Trials, and the Joanna Briggs Institute database. The subject-specific simulation sources searched were the Simulation Innovation Resource Center of the National League of Nurses (SIRC NLN) (U.S.A.) and The International Nursing Association for Clinical Simulation & Learning (INACSL). The Society for Simulation in Healthcare newsletter (SSIH) was searched but not their database of published studies due to the cost and the fact that these articles are available via the databases already searched. Google Scholar and Turning Research into Practice (TRIP) were the general search engines searched for this SR.

The PRISMA Flow Chart for SBCE and Observational Measurement Outcomes SR (Diagram 3) depicts the number of studies retrieved, excluded studies with reasons and the number of included studies. The eight major databases searched retrieved 5716 studies; the simulation specific source studies retrieved 45 studies; and the general search engine search retrieved 27 studies for a total of 72 studies from this search. The total number of retrieved studies was 5788. After removing duplicates from the full set of studies, 3764 studies remained for title and abstract review. The primary and secondary reviewer removed 3729 studies at this point with a 99 per cent agreement rate. The studies that required a consensus decision were reviewed by a third person (thesis committee supervisor) and a decision was reached in this manner.

The remaining 35 studies were reviewed by both reviewers with full text review. Sixteen studies were removed at this point with reasons for exclusion recorded on the study selection database. The two reviewers had an 80 per cent agreement rate at this stage of the review and consensus was reached between the two reviewers, with the thesis supervisor's advice regarding study exclusions due to the population of raters rather than students.

The 16 excluded studies were near-miss study exclusions meaning that these studies would have been included in the SR as they met all inclusion criteria but one. The reasons for their exclusions appear in the PRISMA flow chart (following this section). Nineteen studies remained for inclusion in the SR. These studies included no studies for the years 2000-2005, six studies for the years 2006-2010 and 13 studies for 2011-2015.

Diagram 3. PRISMA Flow Chart for SBCE and Observational Measurement Outcomes SR


Identification
Electronic Databases: CINAHL, Embase, Medline, PsycINFO, Dissertations & Theses (ProQuest), ERIC, Cochrane CENTRAL, JBI
n= 5716
Records identified through other sources: INACSL, SIRC NLN, TRIP, and GOOGLE Scholar
n=72
Search results combined (n=5788)
Screening
Duplicates removed
(n= 2024)
Title/abstract screening
(n=3764)
Articles excluded
(n=3729)
Reasons
Not relevant- (n=3550)
Did not meet criteria-
(n=179)
Eligibility
Full text articles assessed for eligibility (n=35)
Articles excluded (n=16)
Reasons:
Population not UG students (n=8)
Tool not based on observational assessment (n=3)
Tool not used in SBCE (n=2)
Patient not simulator (n=2)
Only abstract available (n=1)
Articles assessed for study quality (n=19)
Included
Articles excluded (n=0)
Reasons
Quality assessment part of SR results
Full text articles included for quantitative synthesis (n=19)

Flow diagram showing article selection according to PRISMA (preferred reporting items for systematic reviews and meta-analyses) criteria (Moher, Liberati, Tetzlaff, & Altman, 2009).

Organization and Presentation of SR Findings

Reporting of SR results begins with a description of the various stages of the selection process which was guided by the PRISMA checklist and flow chart presented in Diagram 3. Next is the report of the search strategy terms guiding the search. The results of the search are depicted in Diagram 3, and in the PRISMA flow chart indicating the search sources, number of articles retrieved, and the number rejected with reasons. The findings of the SR are presented in Appendix F, Tables 3-12 related to each the five objectives of this SR.

Appendix F, Tables 6, 8, and 12 meet recommended reporting of stage one SR synthesis results by grouping studies into logical categories for the SR which are: study design, outcome measurement instrument type, domain of learning, and type of SBCE clinical scenario.

Finally, stage two of the narrative report termed within-study analysis, presents a discussion of quality assessment and findings from each study by discussing whether or not: a) the instrument is clearly linked to concepts in the study; b) the instrument and measures are described objectively; c) the reliability of the instrument is described with supporting statistics; d) the validity of the instrument is described with supporting statistics; and e) whether a detailed protocol is provided for use of the instrument. Appendix F, Table 9 presents the reliability, validity, and quality assessment of the data while the summarization of the studies' COSMIN results are presented in Appendix F, Table 10. The objectives are presented below and reflect the study PICOTS.

Objectives and Findings

Primary objective.

The primary objective of the SR is to identify and assess the characteristics of the outcome measurement instruments used in high fidelity SBCE and the outcomes of competence, CT, CJ, and CR with undergraduate nursing students.

The assessment of the outcome measurement instruments involved examining the collected data from 19 study data extraction forms using the following study qualities: the country of origin of the study; the study source; the study purpose; the study design; the study population (number, characteristics, eligibility, & recruitment); the sample size; the setting; the intervention; the comparative intervention; the outcome measurement instrument; the measurable outcomes of CJ, CT, CR, and competence; the domain of learning captured by the study; all reported statistical analyses; reported instrument reliability and validity; study limitations; and study results.

The researcher also reviewed each study identifying the method of evaluation (rubric, checklist, evaluation tool) and categorized all studies by type of evaluation. The findings for this objective are presented in Appendix F, Table 4 which provides a descriptive summary of study characteristics and in Appendix F, Table 5 which presents a descriptive summary of the instruments' measurable outcomes (CJ, CT, CR, competence). The measurable outcome findings will be presented within the analysis of each study.

As noted in Appendix F, Table 8, six studies employed the Lasater Clinical Judgment Rubric (LCJR) or a modified version of it; five studies used a self-developed rubric; four studies used self-developed checklists; four studies used simulation evaluation tools which included: Clinical Simulation Evaluation Tool (CSET), Creighton Simulation Evaluation Instrument (C-

SEI), Seattle University Simulation Evaluation Tool, and the Heart Failure Simulation Competency Evaluation Tool (HFSCET). In total, there were 11 rubric evaluation instruments, four checklist evaluation instruments and four simulation evaluation instruments.

Objective two.

The second objective is to identify and map the most commonly used outcome measurement tools by type of SBCE scenario and by country of use (i.e. Canada, U.S.A., U.K., and Europe).

Findings for this objective are presented in Appendix F, Table 4 (a descriptive summary of included studies) and Table 6 (the studies categorized by type of SBCE clinical scenario). The majority of the studies in the SR were from the United States (16 of the 19), while one was from South Korea, one from Singapore, and one from the United Kingdom (Table 4). The SBCE scenarios included maternal / child nursing scenarios (two studies), adult medical / surgical scenarios (11 studies), adult cardiology scenarios (eight studies), and others did not define the type of scenario (three studies). Five studies had scenarios in two categories.

Objective three.

Objective three is to characterize outcome measurement tools by the type(s) of learning domain(s) depicted in the SBCE student outcomes.

Appendix F, Table 8 characterizes the studies by domain of learning and the instruments' evaluation methods. Findings related to this objective are also in Appendix F, Table 4 (a descriptive summary of the instruments' measurable outcomes), and Table 5 which presents the learning domains within the descriptive summary of the outcome measurement instruments.

None of the 19 SR studies demonstrated specific performance parameters by which to identify the conative or affective domains. All 19 studies captured the cognitive and psychomotor learning domains.

Objective four.

The fourth objective is to characterize the outcome measurement tools by the outcomes measured in the SBCE.

The SBCE measurable outcomes (CT, CR, CJ and competence) identified from the literature review (Chapter 2, Table 1) provided the list of observable student behaviours against which the researcher compared the outcomes of each study. The findings related to this objective can be found in Appendix F, Table 5 (a descriptive summary of the instruments' measurable outcomes).

In order to determine evidence of SBCE outcomes considered measurable the researcher examined the instrument (where provided) and the reported measurement instrument methodology and outcomes. In the SR literature review the researcher identified commonly measured observable student behaviours and outcomes in SBCE research being used to assess nursing competence, as CT, CJ, and CR (Chapter 2, Table1).

Each of the 19 SR studies was reviewed to identify and record which of the four SBCE outcomes (CT, CR, CJ, and competence) were identified by the author(s). Eighteen studies included CT process outcomes; all 19 studies included CJ process outcomes; 18 studies had CR process outcomes; and 19 studies showed competence process outcomes. The study findings for measurable student behaviour outcomes are presented in Appendix F, Table 5.

According to this SR literature review (Chapter 2, Table 1) of observable student behaviours, critical thinking has four measurable outcomes, clinical reasoning has six measurable outcomes, clinical judgment has eight measurable outcomes, and competence has five measurable outcomes. None of the studies met all the measurable outcomes for all of the SBCE outcomes of CT, CR, CJ, and competence. Four studies met all four measurable outcomes for CT; one study met all eight measurable outcomes for CJ; one study met all six measurable outcomes for CR; and three studies met all five measurable outcomes for competence.

Objective five.

Objective five is to characterize outcome measurement tools based on the methods used to establish their reliability and validity.

Findings related to reliability and validity of outcome measurement tools are found in Appendix F, Table 9 (outcome measurement instruments' reported validity and reliability) and in Appendix F, Table 10 (the COSMIN study/ instrument quality assessment ratings).

The researcher reviewed each study description or reported measures of reliability and validity. These are recorded in Appendix F, Table 9 which also includes reported study statistical analyses, participant eligibility and recruitment, a description of the setting and generalizability findings.

Statistical analyses methods were reported for 15 of the 19 studies. The statistical analyses reported for the 15 studies were: descriptive statistics in six studies; analysis of variance (ANOVA) in six studies; t-tests in six studies; chi square analysis in four studies; Z-test of proportion in one study; Mann Whitney U in two studies; Wilcoxon Signed ranks and Spearman rho in one study; Scheffé post-hoc tests in one study; Kolmogorov-Smirnoff and Shapiro-Wilk statistics in one study; and a power analysis statistic in two studies.

All studies reported participant eligibility and recruitment which was most frequently by convenience sampling as the participating students were enrolled in a course in which the study took place. All the studies took place in a school of nursing simulation lab (intervention groups), classroom, or clinical setting (control groups). Generalizability of study results was limited due to small sample size (n=4), homogeneity of population or lack of demographic data to establish population diversity (n=11), convenience sampling (n=13), studies carried out at only one site (n=9), and other research design or methodology reasons (n=3).

Nine studies reported internal consistency reliability findings using Cronbach's α statistic. Six studies reported equivalence reliability / inter-rater reliability (IRR) using: rater agreement index (one study), dependent sample t- tests (one study), per cent agreement (one study), Cohen's Kappa (two studies), Cronbach's α (one study), and by Pearson product moment coefficient (one study). Three studies had one rater only so did not require IRR. Stability test-retest reliability findings were reported for three studies employing Pearson's correlation coefficient. Stability, internal consistency and equivalence reliability measures were reported by each of three studies. Eight studies did not indicate any reliability measures.

Content validity was reported for 13 of the 19 studies with nine studies reporting a literature review as a method of establishing content validity and six studies reporting use of an expert panel to establish content validity. One of the studies used both literature review and a panel of experts as well as conducting content validity index statistics at both the item and scale level to establish content validity of the instrument.

Criterion validity was reported for two studies: one study where the evaluation instrument was based directly on medication administration rights and for another study where the criterion was based on the Airway Breathing Circulation Disability Exposure (ABCDE) assessment tool. Construct validity was reported for two studies: construct validity was established in one study with a contrasted group procedure and ANOVA f-test statistics whereas the other study employed Cronbach's α and factor analysis statistics. Both content and construct validity were reported for one study.

Only one study reported three methods of instrument reliability and two methods of instrument validity. Seven studies did not report reliability measures but three of these studies did report one method of establishing instrument validity. Three of the studies had no reported reliability or validity measures.

Quality Assessment of SR Studies

In order to assess the quality of each study in the SR, the Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) checklist was chosen as it is particularly useful in SRs to appraise the methodological quality of studies on measurement properties (Mokkink et al., 2010). The COSMIN checklist is comprised of 12 content areas termed *boxes* used to assess whether a study meets the standards for good methodological quality. Two boxes evaluate whether general requirements of a study on measurement properties are met. Nine boxes evaluate the quality of the assessment of different measurement properties concerning: internal consistency, reliability, measurement error, content validity (including face validity), construct validity (subdivided into three boxes, structural validity, hypotheses testing, and cross-cultural validity), criterion validity, responsiveness, and generalizability.

In order to begin the quality assessment phase of the SR, the researcher sent the COSMIN website link to the second reviewer a week before quality assessment began in order for the reviewer to take time to read the checklist background, taxonomy, definitions, procedure and become generally familiar with the checklist. The researcher and second reviewer then met (via telephone) in order to orient the second reviewer more fully by going through the COSMIN checklist page by page, checking to ensure both reviewers interpreted the checklist criteria in the same way, and deciding how to proceed with the checklist in hard copy as an online version is not yet available. Three studies were trialed for COSMIN rater agreement before beginning the SR rating resulting in a 75% inter-rater reliability. Both the reviewer and researcher contacted each other when a question arose about the application of the COSMIN checklist in order to ensure consistency of use of the tool.

Where the reviewer and researcher had a different interpretation of an item on the checklist, the question was referred to a third person (the thesis supervisor) for final interpretation before quality assessment of the studies began. This occurred with one question repeated in several of the checklist boxes. The question was: were there any important flaws in the design or methods of the study? Both reviewers felt the question was open to interpretation and that the checklist criteria captured all major study flaws so were hesitant to apply this question. The issue was taken to the thesis supervisor who agreed that for the following reasons we would not include the question: 1) we didn’t anticipate major flaws in the studies because the procedures used would have captured this; 2) any major issues would be captured by COSMIN checklist questions; and 3) making a decision as to what constitutes other major flaws in each study would become too subjective without specific criteria and require consensus decision each time.

The quality review for this SR applied Boxes A (internal consistency), B (reliability), C (measurement error), D (content validity), and the box for generalizability for the studies. Boxes E (structural validity), F (hypothesis testing), G (cross-cultural validity), H (criterion validity), I (responsiveness) and J (interpretability) were not used in this SR as none of the studies had data applicable to boxes E through H and Boxes I and J were considered not relevant to the type of measurement instruments under review. There is no problem in selecting and using only those boxes that are applicable to the studies under review as the COSMIN checklist is meant to be used in a modular fashion such that it is not necessary to complete the whole checklist when evaluating the quality of a study (Mokkink et al., 2010).

The researcher and the second reviewer applied the COSMIN checklist independently to assess study quality of all 19 SR studies. Reviewer two entered her results on an electronic spreadsheet obtained from the researcher's personal communication with Dr. Terwee (January 12, 2015) and then the researcher entered her results for the 19 studies. The researcher and second reviewer compared and discussed results (by telephone) and reached consensus on ratings that were different. This was done for each checklist box in each study. The final ratings appear in Appendix F Table 10.

The inter-rater reliability for the two reviewers was estimated at 68 % for box A (internal consistency), at 73% for Box B (reliability), at 78% for Box C (measurement error), at 36% for Box D (content validity), and at 57% for the generalizability box. The difference on scoring for Box D and the generalizability box came to light during the consensus discussion of scores. The percent agreement on the scores for the content validity and generalizability boxes was lower because the researcher was scoring missing information stringently while the reviewer (who is a seasoned researcher) was able to see the big picture of the study and considered minor loss of detail less reason to reduce the score. At that point the researcher and reviewer reviewed the criteria for each question and box and consulted the COSMIN manual where there was a difference in opinion and came to consensus on a score for each Box.

Both reviewers independently applied the COSMIN tool to each study and gave a rating for each of the applicable properties / boxes. The items in the COSMIN tool have four response options in order to increase the discriminative ability of the items and to represent excellent, good, fair, or poor methodological quality. The methodological quality score per box is obtained by taking the lowest rating of any item in a box (Mokkink et al., 2010). The COSMIN developers recommend using this scoring system when performing a systematic review of measurement properties. Appendix F, Table 10 presents the ratings for each study.

All 19 studies received a rating of poor for internal consistency due to either small sample size, lack of factor analysis or lack of Cronbach's α statistic. The 19 studies also received a rating of poor for reliability due to small sample size, and/or lack of intraclass correlation coefficient (ICC), Pearson or Spearman coefficients. Sixteen studies received a rating of poor for measurement error also due to sample size, and no percentage of missing items given. The remaining three studies had ratings of fair for measurement error due to lack of clarity on how missing items were handled. Content validity was rated as excellent for 17 of the 19 studies, fair for one study (lack of comprehensive construct), and poor for one study (no assessment of item relevance).

None of the studies were rated for structural validity, hypothesis testing, cross- cultural validity, criterion validity or responsiveness due to lack of information in these areas. Generalizability was rated as poor for ten studies (no report of mean age, gender, diversity or demographics reported and due to sampling method); rated as fair for six studies (where mean age and gender were provided), and good for three studies (where more complete demographics were provided).

Within-Study Analysis

Stage two of reporting SR results is a presentation of findings from each study discussing whether or not: a) the instrument is clearly linked to concepts in the study; b) the instrument and measures are described objectively; c) the reliability and validity of the instrument is described with supporting statistics; and d) whether a detailed protocol is provided for use of the instrument. The analysis of each study is presented below. For complete study results on these subjects see Appendix F for the following tables: Table 4 a descriptive summary of study characteristics; Table 5) a descriptive summary of instruments' measurable outcomes: (CJ, CT, CR, and Competence); and Appendix F, Table 9 for validity and reliability results.

Study 1.

The study by Aronson, Glynn, and Squires (2012) is a non-experimental instrument development study carried out in the U.S.A. to develop a simulation competency package and complete initial psychometric testing of a rating tool designed to assess student nurse competency in responding to a deteriorating patient situation.

The study was clearly linked to the SR concepts of competence, CT, and CR and less so to CJ as evidenced by the presence and number of measurable outcomes in each of these areas (Appendix F, Table 5).

The Heart Failure Simulation Competency Evaluation Tool (HFSCET) captured all four CT behavioural outcomes, 5/6 CR outcomes, and 4 /5 competence outcomes. The tool, however, captured only 4/8 CJ outcomes. A unique feature of the simulation was the “practice out loud” strategy whereby the raters were able to evaluate students’ communication, assessment and CR on the tool. Both cognitive and psychomotor domains of learning were captured by the instrument. Both the instrument and the measures were described objectively and in detail. IRR for Phase 1was 0.73, 0.76, and 0.77 for each of three raters; Phase 2 IRR was 0.83 for two raters. Instrument validity was reported narratively for content validity as a literature review and best practices guideline review. A detailed protocol regarding instrument use was not included in the study.

Study 2.

The authors of study two, Ashcraft et al. (2013) conducted a non-experimental descriptive study in the U.S.A. to describe the process of evaluating senior nursing students in a simulation laboratory using a Modified Lasater Clinical Judgment Rubric instrument.

The study was linked to the SR concepts of: CJ by meeting 5/8 outcome behaviours, competence by meeting all but one behaviour (4/5), CT by meeting 3/4 behaviours and, minimally for the CR concept by meeting only 2/6 behaviours. The tool captured the psychomotor and cognitive domains of learning. The authors provided objective but brief explanations of the learning domains, while descriptions of measures were more detailed. Statistical analyses were reported in the study for internal consistency reliability in both phases (.082 - .927), and content validity via expert panel was identified only for the scenarios not the tool. There was no protocol given for the instrument.

Study 3.

The study by Doolen (2012) is a PhD dissertation from the U.S.A. The non-experimental methodological study sought to test the psychometric properties of a new instrument, the Simulation Thinking Rubric (STR) to assess higher order thinking during high fidelity simulation.

Doolen's study is linked most closely to the SR concept of CJ as it met 6/8 measurable outcomes, then to competence by 4/5 outcomes, next to CR by 3/6 outcomes and to CT as it met 2/4 outcomes. Both the cognitive and psychomotor domains were captured by the STR. The instrument and measures were presented objectively and in detail.

The reported reliability statistics were for stability, internal consistency (Cronbach's α = .74), and IRR via a Pearson product moment coefficient which showed that the correlation between raters was not adequate, revealing that the STR is an unreliable instrument. The author reported established validity for STR for content validity (panel of experts) and via content validity index statistics for item (average of .928) and subscale (average of .976). Construct validity was reported via a contrasted group procedure with an analysis of variance (ANOVA) F test which measured the difference between the means of the two groups. Although the author states the STR is a criterion referenced tool (based on the Simulation Based on Learning Language model (SIMBaLL) using Piaget's four cognitive developmental stages) there was no discussion of it in the results. A detailed protocol for the instrument was not included in the study.

Study 4.

Gantt's (2010) quasi-experimental pilot study was conducted in the U.S.A. to apply the Clark Simulation Evaluation Rubric with undergraduate nursing students of different levels from two different types of programs (associate (AD) and baccalaureate (BN) degree) in simulated clinical scenarios.

The study was clearly linked to the SR outcome concepts and most closely to CT as it portrayed all four measurable outcomes, then to competence by 4/5 outcomes, and finally to CR as it met 4/6 outcomes but did not portray any outcomes for CJ. The cognitive and psychomotor domains were captured by this instrument. The study stated reliability findings from the original rubric development studies (Clark, 2007a, 2007b) but did not attempt to establish current IRR or other reliability statistics which the author reported as a limitation of the study. Content validity was reported from the original studies only. There was no protocol for the instrument provided.

Study 5.

The authors of study five, Goodstone and Goodstone (2013) conducted a quasi-experimental pilot (posttest only) study in the U.S.A. to describe the use of human patient simulation to develop a performance-based competency measure of medication administration safety.

The Medication Administration Safety Assessment Tool (MASAT) was linked most closely to the SR concept of competence as exhibited by meeting 3/5 competence outcomes, and only one in each of CT, CJ, and CR (which related to technical skills, safety or reviewing information). Both the cognitive and psychomotor domains of learning were captured by the tool. The authors reported reliability measures of: internal consistency reliability (Cronbach's α .84); and equivalence reliability IRR via rater-agreement index (RAI across 4 raters) at .83 for the 14 student videos and .90 for scores on the pre-recorded student performance examples. Content validity was established via literature review, and by 10 subject matter experts (SME) who indicated strong support for each behaviour as representative of targeted content domain. Content validity was also measured by content validity index (CVI) for both item (I-CVI: three=0.75, one = .88, rest =1.00) and subscale (S-CVI: .93) which exceed standards considered acceptable. The MASAT is a criterion referenced measure via the six rights of medication administration and safety. The tool and the measures were described objectively and clearly. A brief protocol for use of the tool was provided

Study 6.

Haggard's (2013) PhD dissertation was a non-experimental correlational design study conducted in the U.S.A. to investigate whether high-fidelity simulation scenarios fostered nursing student safety competencies in the provision of nursing care as measured by the Sweeney-Clark Simulation Performance Rubric: with Haggard Modification.

The study was linked most closely to the SR concepts of CT (3/4 outcomes met), competence (3/5 outcomes met) , then to CJ (4/8 outcomes) and lastly to CR (2/6 outcomes met). Both the cognitive and psychomotor domains of learning were captured by the tool and both the tool and measures were described objectively and in detail. The author reported reliability statistics for stability reliability (Pearson's coefficient), for internal consistency reliability (Cronbach's α = .86 to .96 from original tool/study and 0.78-0.93 for current study), and for equivalence reliability (IRR by dependent sample t-tests which were non-significant indicating a good level of agreement). The author states that content validity was achieved by establishing IRR with the instrument. Construct validity was determined with Cronbach's α (.80) and factor analysis. A brief protocol for use of the instrument was provided.

Study 7.

Jensen (2013) conducted a quasi-experimental descriptive study in the U.S.A. with the purpose of using patient simulation in associate and baccalaureate nursing degree capstone courses to evaluate nursing students' clinical reasoning skills during patient simulation using the Lasater's Clinical Judgment Rubric (LCJR) and to compare students' self-assessed and faculty assessed ratings of clinical reasoning skills.

The LCJR in this study was most closely linked to the SR concept of CJ (5/8 outcomes), then to CR (3/6 outcomes), next to competence (2/5 outcomes), and not at all to CT (0 outcomes). Study tool and measures were described objectively. The LCJR captured both cognitive and psychomotor domains of learning. Internal consistency reliability measures included Cronbach α for the entire LCJR scale (α = .95) and for each subscale: noticing (α =

.88); interpreting (α = 0.88); responding (α = .88); and reflecting (α = .86). Validity measures for the study were not described, nor was a protocol for use of the instrument.

Study 8.

Kim and Shin's (2013) quasi-experimental pretest-posttest study set in South Korea aimed to develop prenatal, labor/delivery and postpartum clinical simulation scenarios for obstetrical nursing students and to provide educational tools for student evaluation.

The unnamed 15 item checklist developed to evaluate the students' skills and attitudes in each of the different scenarios was linked most closely to the SR concept of CT (met 4/4 outcomes), then to competence (3/5 outcomes) and, minimally to CR (1/6 outcomes) and CJ (1/8 outcomes). Both the cognitive and psychomotor domains were captured by the tool. The authors described the tool and measures objectively. No reliability or validity methods or statistics were provided nor was a detailed protocol for use of the instrument.

Study 9.

An exploratory mixed methods (qualitative-quantitative-qualitative) study by Lasater (2007) in the U.S.A. aimed to describe students' responses to simulated scenarios within the framework of Tanner's (2006) Clinical Judgment Model and to develop a rubric that described levels of performance in clinical judgment. The resulting rubric was titled the Lasater Clinical Judgment Rubric (LCJR) and has either been used, or adapted for use in numerous other studies.

The LCJR used in this study was linked most closely to the SR concept of CJ (7 /8 outcomes), then to CT (3/4 outcomes met), next to CR (4/6), and lastly to competence with 3/5 outcomes met. According to the tool and the author's description all four learning domains were captured by the LCJR. No reliability methods or statistics were reported however the content validity was established via a tool development cycle of description-observation-revision-review by an expert in CJ, and an expert in rubric development and repeated weekly until testing time. The tool and measures were described objectively, however, no protocol for tool use was included.

Study 10.

Lasater's 2005 PhD dissertation was conducted in the USA to evaluate the impact of high fidelity simulation in the development of clinical judgment. The exploratory design used both qualitative and quantitative methods and further development of the LCJR. This version of the tool was titled the Lasater Clinical Judgment in Simulation Rubric (LCJSR).

The simulation rubric employed in this study was linked to the SR concept of competence (4/5), next to CT (3/4), then to CJ (6/8 outcomes), and last to CR (3/6). The tool captured the cognitive and psychomotor domains. No reliability statistics were provided, however, the author noted that internal consistency was poor as the rubric was continuously refined during the scoring phase, and thus neither the rubric nor the actual scores were consistent.

Content validity was established by modeling the tool on Tanner's (2006) Model of Clinical Judgment and by collaboration with an expert in educational rubric development with tool revision after observing students in simulation. Both the tool and the measures were described objectively with detail but no protocol for tool use was provided.

Study 11.

Liaw et al. (2010) conducted a quasi-experimental cross-over intervention posttest study in Singapore to evaluate the integration of a simulation based learning activity on nursing students' clinical crisis management performance in a problem-based learning (PBL) curriculum.

Liaw et al. developed two unnamed sets of checklists that were linked to the SR concepts of CT (3/4 outcomes), to CR (4/6 outcomes), to and CJ (3/5 outcomes), and least to CJ at 2/8 outcomes. Cognitive and psychomotor domains of learning were captured by the tool. No reliability statistics were given and IRR was not established as there was only one marker. Content validity was established by a panel of nursing and medical experts' review of the checklists which were revised after pilot testing with two students. The tool and measures were clearly and objectively described however the authors did not present a protocol for use of the instrument.

Study 12.

Merriman et al. (2014) carried out an experimental randomized, controlled trial study in the UK to determine whether clinical simulation is more effective than traditional classroom teaching in teaching the assessment skills required to recognize an acutely unwell, deteriorating patient. The tool used to evaluate student performance on an HPS during an OSCE was an unnamed Checklist of 24 performance criteria.

The tool was most closely linked to the SR concept of CJ as it met 5/8 outcomes, then to competence at 3/5 outcomes, next to CR at 3/6 outcomes and to CT at 2/4 outcomes. The checklist captures the cognitive and psychomotor domains of learning. The authors did not report content validity measures, however the checklist was based on the Airway Breathing Circulation Disability Exposure (ABCDE) assessment tool which was described. No reliability measures were reported. The tool and measures were described objectively but no protocol for use of the checklist was reported.

Study 13.

Meyer's (2012) PhD dissertation study conducted in the U.S.A. is a quasi-experimental non-randomized controlled trial to explore the effects of using simulation and didactic instruction on students' critical thinking and clinical judgment. The tool is a modified version of the Lasater

Clinical Judgment Rubric referred to in the study as the Modified LCJR, where the modifications to the LCJR involved the elimination of the planning phase of the nursing process and having each scenario rubric designed around interventions and procedures for that particular scenario.

The Modified LCJR was most closely linked to the SR concept of competence as it met all five outcomes, next to CT (3/4 outcomes), then to CJ (5/8 outcomes), and last to CR (3/6 outcomes) while capturing the affective, cognitive and psychomotor domains of learning. Neither reliability nor validity measures were reported. The tool was described objectively but the author did not include a protocol for its use.

Study 14.

The U.S.A. was the site for the non-experimental instrument development study by Mikasa, Cicero, and Adamson (2013). The study goal was to create an evaluation rubric for simulated clinical that integrated course objectives with the 1999 American Association Colleges of Nursing (AACN) baccalaureate competencies and provide objective outcome data. The tool is The Seattle University Simulation Evaluation Tool.

The tool was linked most closely to the SR concepts of CT (3/4 outcomes), to CR (5/6 outcomes), then to competence (3/5), and least to the CJ concept (3/8). The cognitive and psychomotor domains were captured by the tool. Reliability statistics for the tool were not measured in the study but the authors reported internal consistency (Cronbach's α = .97), IRR via intra-class correlation (0.85), and test-retest reliability using intra-class correlation (0.90) from the tool results in the study by Adamson and Kardong-Edgren (2012). Content validity was established through measures taken to base the tool on the AACN's (1999) core competencies for baccalaureate nursing education and clinical course objectives at the study university. An expert in evaluation scales was consulted during tool development and revision. The tool and measures were described objectively; no protocol for tool implementation was included.

Study 15.

Nicholson's (2010) PhD dissertation is experimental posttest-only study carried out in the U.S.A. to determine if there were differences in active learning teaching strategies (case-based learning, simulation with narrative pedagogy, and simulation) on the outcomes of nursing student performance of intervention activities, performance retention of intervention activities, student satisfaction, self-confidence, and students' educational practice perceptions.

The Student Performance Demonstration Rubric was linked equally to the SR concepts of CR and competence as it met all outcomes in each concept, then CT (3/4 outcomes met), and lastly to CJ (5/8 outcomes met). The rubric captured the cognitive and psychomotor domains of learning. Internal consistency reliability was reported (Cronbach's α = .92) and IRR at 0.92 (one team member scored the rubric while viewing the recorded demonstration and a second scored the rubric every fifth recording). Content validity measures included an expert panel of nurse educators and basing the rubric on American Heart Association guidelines for care of patients with myocardial infarction. The tool and measures were reported objectively and clearly yet no protocol for use of the rubric was included.

Study 16.

Patton's (2013) non-experimental descriptive study carried out in the U.S.A. sought to assess the degree of agreement between the ratings of student performance during a clinical simulation by critical care course instructors and the course coordinator using The Creighton Simulation Evaluation Instrument (C-SEI).

The C-SEI was linked most closely to the SR concepts of CT (3/4 outcomes), and competence (2/5 outcomes), less directly to CR (2/6 outcomes), least to CJ (2/8 outcomes), and captured the cognitive and psychomotor domains of learning. Inter-rater reliability IRR between two raters was 0.85 to 0.89 per cent agreement for the categories assessment, communication, critical thinking, and technical skills. Content validity was established through a literature review and expert panel however no further details were provided. The study tool and measures were described briefly but objectively and did not include a protocol for use of the tool.

Study 17.

Radhakrishnan, Roche, and Cunningham (2007) conducted a quasi-experimental two group posttest pilot study in the U.S.A. to identify nursing clinical practice parameters influenced by simulation practice and to measure clinical performance improvement. The Clinical Simulation Evaluation Tool (CSET) was related to the SR concepts of competence (4/5 outcomes), CT (3/4 outcomes), to CR (4/6 outcomes), and last to CJ (4/8 outcomes). The tool reflected the cognitive and psychomotor learning domains. The authors presented the tool objectively but did not report any reliability or validity methods or statistics nor was tool protocol included.

Study 18.

Strickland's (2013) experimental randomized pretest/post-test study is described in the author's PhD dissertation conducted in the U.S.A. Strickland compared the accuracy of student's self-assessment of clinical judgment skills with faculty assessment of the student's clinical judgment skills upon completion of a high-fidelity simulation experience. As well, the study aimed to: compare the relationship between student's self-assessment and faculty assessment of clinical judgment competency levels during HPS, and student's scores on a customized HESI nursing exam; and to examine how high-fidelity simulation influences nursing student's clinical judgment competency level. Lastly, the author examined how students who experience a simulated clinical event perform on the content specific HESI (post intervention) versus those who do not receive the simulated clinical.

Strickland used the Lasater Clinical Judgment Rubric (LCJR) which, in this study, aligned perfectly with CJ concepts (8/8 outcomes), competence (5/5 outcomes), closely to CT (3/4outcomes) and CR (5/6 outcomes). The rubric captured the cognitive and psychomotor learning domains. Reliability measures included internal consistency (Cronbach's α = .82), and stability reliability with Pearson's correlation coefficient statistic (r = .314). Equivalence reliability was not reported as only the author rated the students using the LCJR. Content validity methods were not reported for this study. The tool and measures were described objectively and clearly; an instrument protocol was not provided.

Study 19.

Swanson et al. (2011) carried out an experimental posttest-only study in the U.S.A. to compare the effects of three active learning strategies on the outcomes of intervention activities, performance retention of intervention activities, student self-confidence, and student educational practice preferences.

The Student Performance Demonstration Rubric was most closely linked to the SR concepts of competence (3/5 outcomes), and CR (3/6 outcomes), then to CT (2/4 outcomes), least to CJ (2/8 outcomes), and captured the cognitive and psychomotor domains of learning. Student performance videos were viewed and scored by one researcher and every fifth recording was also scored by a second researcher with IRR rater agreement of 0 .90 for the first performance and 0.94 for the retention performance scores. Content validity included basing the rubric on the American Heart Association myocardial infarction guidelines. The rubric was described objectively without a protocol provided for its implementation.

Summary

The SR search provided nineteen eligible studies representing 16 different simulation outcome measurement instruments which were predominantly from the U.S.A. The instruments exclusively captured the cognitive and psychomotor learning domains with none capturing the affective or conative learning domains, with over half being rubric design tools. The studies were clearly linked to the concepts of competence, critical thinking, clinical judgment, and clinical reasoning though at varying levels for each concept. Clinical simulation scenarios in the studies primarily involved adult medical surgical events with only two maternal/child scenarios and three unidentified.

The methods used to establish reliability of the outcome measurement instruments were most frequently Cronbach's alpha, per cent agreement for IRR, and Cohen's Kappa. Content validity was most frequently established by literature review and expert panel review. The results of the COSMIN quality assessment review indicated that: a) all 19 studies were lacking in reliability measures but strong in content validity measures; b) 16 studies had a rating of poor for measurement error; and c) over half the studies had a poor rating for generalizability. Sample size, sampling methods, and lack of reported information were factors in many areas. The poor COSMIN reliability study ratings appear to be in contradiction with the acceptable statistical reliability findings for many of the individual studies, however, the COSMIN criteria for reliability are inclusive of many factors (described in the discussion chapter). Any one rating of poor for any of these criteria results in a rating of poor for study reliability as a whole. Therefore, although the study meets some reliability standards if it does not meet COSMIN standards the study receives a rating of poor with this tool.

The within-study results provided a synopsis of each study as per the study objectives. Chapter 5 will present a cross-study synthesis of the study findings taking into consideration variations in study quality and other variables that may affect generalizability of studies.

Chapter 5 Discussion

Introduction

This study focused on competence measurement with undergraduate nursing students and three issues related to it: (a) the lack of a consistent definition of competence and related outcomes, (b) varying evidence on the ability of simulation-based clinical experience (SBCE) to measure competence and, (c) a lack of clarity on which observational measurement tools (i.e. in current use with SBCE and high fidelity patient simulators) are a valid and reliable measure of undergraduate students' individual competency outcomes.

The aim of the proposed systematic review (SR) was to determine which observational outcome measurement (evaluation) tools provide measures that are both valid and reliable in measuring undergraduate nursing students' outcomes of nursing: 1) competence, 2) clinical judgment (CJ), 3) clinical reasoning (CR), and 4) critical thinking (CT) following a high-fidelity simulation-based clinical experience (SBCE). In addition, the purpose was to identify studies that quantitatively identified and measured these outcomes. The SR will also provide an analysis of the relationships between the SR studies, an overall summary discussion of study findings and an assessment of the robustness of the evidence as is expected of a SR (CRD, 2009). The strengths and limitations for this SR, gaps in current SBCE observational measurement tool knowledge and recommendations for future research follow next. Concluding the chapter is a discussion of SR results as they impact key stakeholders such as simulation users and educators, researchers as well as patients and health care providers.

Identifying Concepts in SBCE Measurement Research

In order to develop or select an existing instrument to measure the student outcomes of a SBCE, the researcher or simulation educator must first identify what concept(s) they are seeking to capture and ultimately, measure. Once the concept is clearly identified and an operational definition is established then it is possible to accurately identify performance behaviours that are both observable and measurable depictions of the desired student outcome(s). The clearly defined concept, operational definition, and behavioural descriptors create a tool that will enable the observer or rater to determine when the student has met (or not met) the desired behaviours and outcomes.

However, it was clear from this SR literature review, that there continues to be a lack of consensus on the definition and meaning of the concepts and use of the terms competence, critical thinking (CT), clinical judgment (CJ), and clinical reasoning (CR) as they relate to the profession of nursing.

While the term competence would seem to be a word with one accepted meaning, its meaning is often confused with the term competency in the nursing literature where they are used differently with inherently different meanings. The differing views about the definitions of the words competence and competency are: that competence is an aspect of a job that a person performs while competency is the behaviour underpinning the performance (Cowan et al, 2007); or that competence is the knowledge, capacity and potential to perform skills while competency is the actual performance in accordance with policies in a situation (Cowan et al.). Where the confusion often lies is determining whether one term expresses knowledge and capability while the other term expresses performance, or essentially, do they mean the same thing?

Benner's (1982) definition of nursing competence as "the ability to perform a task with desirable outcomes under the varied circumstances of the real world" incorporates both capability (ability) and performance (to perform a task) in one term and this definition is frequently quoted in nursing literature (Cowan et al.). However, when discussing the three concepts of competence: behaviourist, generic and holistic (Cowan et al., 2007; Garside & Nhemachena, 2013) the meaning of the terms competence and competency again become unclear. The behaviorist or performance concept of competence (as defined by Gonczi, 1994 in Cowan et al.) is task-based, making the task synonymous with performance competency; the generic concept of competence is described as person-oriented including the underlying characteristics and qualities of the individual as indicators of effective performance (Cowan et al., 2007; Fahy et al., 2011) aligning this definition more closely with the ability aspect of competence, while the holistic integrated concept of competence is inclusive of the general attributes of the nurse and the practice context drawing on the nurse's knowledge, skills, attitudes, values and professional judgment for effective performance (Fahy et al., 2011). This definition includes both the practitioner's capacity (capability) and their ability to integrate them in their practice (performance) and is inclusive of both the behaviourist and generic concepts. The mixed use of the terms competence and competency can lead to confusion in the SBCE literature and in particular to outcomes associated with the SBCE, where it is unclear which meaning the researcher has assigned to the term competence thus inhibiting a mutual understanding of the reported outcomes.

The International Nursing Association for Clinical Simulation and Learning (INACSL) provides standardized definitions of simulation terms (including definitions for competence, CJ, CT, and CR) presented in The Standards of Best Practice Standard 1: Terminology (Meakim et al, 2013). They are presented with the intent to "promote consistency and understanding in education, practice, research, and publication" (p. S4) among those involved in simulation-based experiences. The SR findings indicated, however, that few researchers made use of these standard definitions.

Furthermore, there is a lack of clarity surrounding the relationship of competence, CR, CJ, and CT to each other, and to the outcome of competent practice. For example, some nursing studies measure CR as a final outcome of the SBCE performance while other studies include CR as a measurable behaviour leading to a final outcome of CJ or competence. In other cases a study includes CT, CR, and CJ as measurable performance behaviours leading to a final outcome of student competence. All three concepts are considered critical to nursing competence, but it remains unclear if CT, CR and CJ are part of the competence process or individual outcomes of their own.

This lack of clarity then impacts measurement strategies and instruments designed to capture each concept. For example, should each concept be measured individually or together? Can they be captured in one tool or do they each require a separate tool? Can competence be assessed by focusing on individual competencies in light of the interaction between competencies? Nursing scholars, researchers, regulators, and practitioners continue to debate on how to define each of these concepts, and whether one concept is more or less essential than the others for the practice of nursing. However, the lack of consensus in definitions for these key concepts is impacting research in competence measurement.

Evidence Regarding use of Definitions for Competence, CT, CR, and CJ

Review of the 19 studies in this SR showed that only three of the researchers defined the competence-related terms (Lasater, 2007; Meyer, 2012; Strickland, 2013), while other researchers either did not use any of the competence-related terms (Kim & Shin, 2013; Merriman, Stayt, & Ricketts, 2014; Nicholson, 2010; Radhakrishnan, Roche & Cunningham, 2007; Swanson et al., 2011), or if the researcher (s) did use these terms they did not provide an operational definition to accompany them (Ashcraft et al., 2013; Aronson, Glynn, & Squires, 2012; Gantt, 2010; Goodstone & Goodstone, 2013; Jensen, 2013; Haggard, 2013; Liaw et al., 2010; Mikasa, Cicero, & Adamson, 2013; Patton, 2013). Instead, researchers often provided a background discussion of one or more of the competence-related concepts in their literature review. For example, Jensen's (2013) study sought to test student's CR with the Lasater Clinical Judgment Rubric, and while she did define CR she did not define CJ. In place of an operational definition of CJ, Jensen provided information on the rubric's foundation in Tanner's (2006) Clinical Judgment Model and identified the four major constructs of CJ. This background helps the reader understand the concept of CJ, but does not clarify how the researcher is applying the constructs to the CJ variable.

The lack of an operational definition for the outcome term used in the study is an unexpected finding as it was anticipated that all studies would include an operational definition of the specific competence-related term(s) for the study. It was, however, anticipated that the definitions and performance behaviour descriptors would vary according to the SBCE researcher's chosen definition and SBCE objectives. Identifying an operational definition is particularly important as the provision of an operational definition for each variable in the research study is an expectation / requirement of quantitative studies; this definition indicates how the variables will be observed and measured in the study (Polit & Beck, 2008). Without this definition it is not clear what performance behaviours or measures will accurately capture the underlying construct of each variable thus decreasing study validity (Polit & Beck). As well, SBCE literature shows that it is critical that an outcome is clearly defined in order for the performance measurement instrument to succeed in capturing the intended outcome (Kardong-Edgren et al., 2010). The inconsistent use of competence-related terms was discovered not only in the literature review but was also a complicating factor during the SR study search; the terms

CT, CR, CJ, and competence located studies with one or more of the terms in the title, but upon review of the abstract, the researcher's application of the term was not found.

It was also expected that the term competence or competency would be used in studies measuring student performance outcomes following a SBCE, however, these terms were only used in eight of the 19 studies (Aronson, Glynn, & Squires, 2012; Gantt, 2010; Goodstone & Goodstone, 2013; Haggard, 2013; Liaw et al., 2010; Mikasa, Cicero, & Adamson, 2013; Patton, 2013; Strickland, 2013). Competence was qualified as safety competence, clinical competence, core competencies, nursing student competence, student patient competencies, and competent performance. These qualifying terms appear to be an attempt to provide clarity on the meaning of the term competence but does not provide enough detail to assist the reader to determine whether the study is addressing the concept with the meaning the reader is seeking. Without a defined conceptual basis and the related outcomes expected from the SBCE, the SBCE researcher or educator is hampered in their ability to focus the goal of the SBCE on improved or changed knowledge, skills, and attitudes of the learner.

The lack of use and definition of the term competence can be related to the issues surrounding the lack of consensus on the concept and definition of nursing competence especially as it relates to competence measurement in nursing education and research. As presented in the literature review, these issues include the confusion surrounding the terms competence and competency, political factors within nursing education, service, and regulation as well as scholars differing viewpoints on the concepts of competence. The behaviorist concept focuses on the nurse's skill and task-based behaviours (e.g. technical skills, knowledge of facts), and their associated level of performance. The generic concept is focused more on the person as nurse including the underlying qualities and personal characteristics (e.g. critical thinking capacity) as indicators of effective performance. The holistic concept definition takes into consideration the general attributes of the nurse and the practice context, which draws on the nurse's knowledge, skills, attitudes, values, and professional judgment (Fahy et al., 2011) thereby incorporating features of both the behaviorist and generic concepts.

Cowan et al. (2007) believe that an acceptance of the encompassing holistic concept of nursing would facilitate acceptance of the competence concept, and provide researchers with a definition to guide research and establish competence standards. This is a feasible suggestion as the SR results show that many of the SBCE researchers are already using SBCE performance indicators and outcomes that are represented in the holistic concept of competence.

SBCE Outcome Competence Indicators and Terminology

Nursing researchers generally seem reluctant to select and define competence-related terms but are more specific in stating the expected student performance behaviours/indicators required to reach the desired competence, CR, CT, or CJ outcomes. The terms used to measure students' competent performance outcomes in the included studies on SBCE were varied and included: (a) *patient safety behaviours*, (Aronson et al., 2012; Goodstone & Goodstone, 2013; Haggard, 2013), (b) *communication skills* (Gantt, 2010; Liaw et al., 2010; Mikasa, Cicero, & Adamson, 2013; Patton, 2013; & Strickland, 2013), (c*) assessment skills* (Aronson et al., 2012; Gantt, 2010; Goodstone & Goodstone, 2013; Haggard, 2013; Liaw et al., 2010; Mikasa, Cicero, & Adamson, 2013; Patton, 2013; & Strickland, 2013), (d) *interventions / patient intervention care* (Aronson et al., 2012; Haggard, 2013; Mikasa, Cicero, & Adamson, 2013), (e) *technical skills* (Patton, 2013), (f) *documentation skills* (Aronson et al., 2012), (g) *patient teaching* (Gantt, 2010; Liaw et al., 2010; & Strickland, 2013), (h) *recognizing the need for diagnostic tests* (,Gantt, 2010;), (i) *professionalism* (Mikasa et al., 2013), (j) *critical thinking* (Gantt, 2010;

Haggard, 2013; Mikasa et al., 2013; Patton, 2013), (k) *teamwork* (Haggard, 2013), (l) *information gathering* (Haggard, 2013), and (m) *informatics* (Haggard, 2013).

The above behaviours/indicators reflect the knowledge (i.e. recognizing need for tests & informatics), skills (i.e. communication, assessment, and intervention skills), attitudes, and values (i.e. teamwork), as well as professional judgment (i.e. CT, professionalism) included in the holistic view of competence. These descriptive indicators then become a quasi-operational definition of competence as they are applied in each study, meaning that even though the researcher did not select or define an operational definition for competence, the researchers are actually aligning themselves with one of the concepts of competence by the indicators they choose.

It is clear from the SR studies that certain indicators for measuring student competence reoccur in five domains (communication skills, assessment skills, intervention skills, CT, and safety skills). Therefore these five domains of skills could be considered common indicators for nursing competence following SBCE.

Accepting these five indicators as necessary domains for all SBCE outcome measurement would be a starting point in providing clarity in competence measurement of undergraduate nursing students and could be a step toward consensus on a definition of nursing competence. Each of the five domains are represented in the holistic concept of competence: holistic competence includes the general attributes of the nurse (communication skills as pertains to the person's ability to clearly and appropriately interact with patients), and the practice context (assessment of environment as well as patient) which draws on the nurse's knowledge, (assessment, patient, safety, and intervention knowledge), skills (assessment skills, safety skills, intervention skills) , attitudes, values, and professional judgment (CT, communicating value for the patient, portraying a caring and professional attitude). Garside and Nhemachena (2013) contend that the holistic definition also provides a basis for transferable skills in delivering care and tools by which to measure it.

Clinical judgment indicators and terminology.

The term clinical judgment (CJ) was used in five of the 19 studies (Ashcraft et al., 2013; Lasater, 2005 & 2007; Meyer, 2012; Strickland, 2013) and phrased as CJ competency, CJ skill, and skill or competency in using CJ. While Strickland provided an operational definition of CJ she also noted that there are conflicting definitions for the term and that they are also used synonymously. The performance behaviours related to the student CJ outcomes ranged from prioritizing nursing interventions, employing evidence-based skills, communication skills, and evaluation of interventions (Ashcraft) to behaviours related to noticing, interpreting, responding, and reflecting (Lasater 2005 & 2007), and student behaviours related to considering conflicting, complex factors and choosing the best course of action for multiple patients (Strickland). Meyer's measurable behaviours for CJ were captured in four scales: assessment, diagnosis, interventions, and evaluation.

While each researcher is actually measuring similar outcomes, often with the same instrument, each study uses different terms and measurable performance behaviours (indicators) which continues to make it difficult for SBCE educators and researchers to compare studies and measurement instruments as they are not based on a shared understanding of the meaning and application of the term CJ. Furthermore, the use of variable terminology for the same indicator adds confusion to SBCE practice, education, and research efforts thus inhibiting mutual understanding and, inevitably, the development of accurate measurement instruments for use in SBCE research and education. The development of evaluation measures in evidence-based practice is critical because psychometrically sound measurement instruments allow researchers to conduct rigorous research regarding current and future educational practices in SBCE which will then allow nursing scholars to make decisions on methods used in innovation and improvement in clinical simulation teaching and learning (Adamson & Kardong-Edgren, 2011).

Currently, nursing students integrate nursing theory and practice in their clinical practice where instructors evaluate student nurse competence through brief periods of observation of students' performance and decision making skills as the students provide patient care (Cowan, Norman, & Coopamah, 2007). It is not possible to ensure clinical placements or patient assignments that ensure each student will receive specific clinical experiences necessary to prepare them to enter complex care environments (Isaacson & Stacy, 2009). Therefore, the creation and implementation of simulation methods providing students with the opportunity for consistent learning and testing environments, complex patient assignments, and evaluators who are skilled in the evaluation method and instruments (Brewer, 2011; Buykx et al., 2011) are needed due to the shortage of clinical sites. Simulation will also provide a venue to improve the performance of health care professionals, reduce human error, and increase patient safety (Canadian Patent Safety Institute, 2008).

A tool that measures nursing student level of competency can demonstrate the improvement in performance and knowledge application during patient care simulation and can therefore be beneficial to students, faculty, and programs in identifying clinical progression standards. Furthermore, the SBCE student competence outcomes can also reflect patient outcomes related to patient safety in the clinical area (Haggard, 2013).

Critical thinking indicators and terminology.

Four studies used the term critical thinking as part of the study: Gantt (2010), Meyer (2012), Nicholson (2010), and Strickland (2013). Nicholson focused on CT as a measurable student outcome but described it broadly as the student's ability to perform nursing interventions which then demonstrates their ability to think critically and apply nursing knowledge. Gantt however, was precise in describing and leveling CT outcomes from low to high (beginning level CT was expressed as the student verbalizing the norms in the patient's condition, while the highest level of CT was reached with the student being able to discuss a plan to avoid patient complications). Strickland (2013) discusses CT as a building block to CJ, likewise, Meyer (2012) describes CJ as an outcome of CT, CR, and the nursing process.

These examples provide evidence that descriptors for CT related terms range from vague to specific with little to no overlap of terms, thus continuing the debate over how to define and measure CT in education and, in particular, nursing education. The lack of consistency in defining and measuring CT was expected and consistent with the SBCE and CT literature, showing that there continues to be a lack of consensus in defining CT, CJ, and CR in nursing. There is agreement in the nursing literature that CT is a core competency for the professional nurse (Ravert, 2008; Smith, 2012) and therefore, it is critical to nursing education and practice that consensus be reached on a definition for CT. Until consensus is reached on a CT definition, the confusion between CT, CR, and CJ will continue to create unnecessary confusion for researchers and readers of SBCE outcomes and educators seeking to select an instrument that captures CT.

Clinical reasoning indicators and terminology.

Clinical reasoning was the least used term in the 19 studies. Two studies measured student CR abilities but neither offered a definition. Jensen (2013) discussed CR competence by measuring students' CR skills as they made a clinical judgment measured by the Lasater Clinical Judgment Rubric (LCJR) via the CJ dimensions of the tool. Likewise, Strickland (2013) used the LCJR to determine students' clinical judgment but stressed that CR is a practice-based form of reasoning and part of CJ. These studies offer similar views on the connection between CR and CJ and use the same tool to measure their outcomes, however, Jensen measured CR competency outcomes while Strickland measured CJ competency outcomes. The ability to apply the same measurement instrument to measure both CR and CJ reinforces the interconnected nature of the two concepts but continues to blur the line between the definitions of CR and CJ, as well as CT and CJ.

The literature provides evidence of the inter-related nature of CT, CJ, CR, and competence, and identifies each as a process leading to the outcome of competent practice. Yet, there continue to be inconsistencies in reporting the relationship of the terms to each other, and to the outcome of competent practice in SBCE research. It is the opinion of this researcher that the lack of clarity surrounding the meaning of the competence-related terms hinders SBCE educators and researchers in the identification and use of the most appropriate language and measurement instrument to match their desired SBCE student outcomes.

In nursing education the problem goes beyond SBCE practice and research as nursing educators use the terms CT, CJ, and CR (often interchangeably) in describing baccalaureate curriculum and course objectives, in course-related literature and in clinical evaluation tools, trusting that their meaning is clearly understood. Unless there are ongoing faculty discussions about program-approved definitions and use of the terms CT, CJ, and CR and their intended application, educators may interpret and integrate them differently in their individual courses. For example, when a clinical simulation lab instructor shares with a student that they are weak in the area of clinical reasoning and uses specific examples and outcome expectations from the SBCE the student can clearly see where they need to improve; conversely, a clinical site instructor may share the same student weakness with the student but base it on the evaluation tool definition and the instructor's interpretation of its meaning. When nurse educators cannot agree on a shared meaning and use of these terms it is not surprising that students have difficulty in understanding what we expect of them and how they can develop proficiency in this area. It is imperative that the concepts of CT, CJ, and CR are systematically defined and used within the curriculum of nursing programs in order to provide clarity and mutual understanding.

Describing student success related to competence, CT, CR, and CJ

The findings of this study show that researchers use various terminologies to describe whether the student met the expected performance outcomes of the SBCE. Student performance was rarely stated as competent except in the following studies: Goodstone and Goodstone (2013) used the term competent for a passing student performance but used the term below standard performance for a failing performance, while Haggard (2013) described levels of student competency, and Aronson et al. (2012) rated the student performance as competent or not competent. Other methods of describing student performance include the terms: accurate or inaccurate performance of essential care (Nicholson, 2010; Swanson et al, 2011), stating that the student accomplished the expected level or not (Jensen, 2013), and whether the student demonstrated CJ skills or CJ competence (Lasater, 2005). In studies where none of the competence-related terms were used to describe SBCE outcomes, researchers used more general terms to describe expected student performance results such as performed actions, capable or not capable of performing the skill measured by the correctness of the action (Kim & Shin, 2013), clinical performance (Radhakrishnan et al., 2007), student performance (Merriman et al., 2014), and student performance of intervention activities (Nicholson, 2010; Swanson et al., 2011).

It seems that nursing researchers and educators are hesitant to identify student performance success or failure as either competent or not competent even when measurement instruments are designed for exactly that purpose. It is probable that one reason for this is the lack of consensus in the literature and nursing in general, not only on the definitions of competence, CT, CR, and CJ but also about what level of competence is acceptable for students in different program years, and for the new graduate nurse. Educators do have specific learning objectives to prepare their students to meet course outcomes and become competent at that level in order to move on to the next level of learning. As well, regulatory bodies describe entry level competencies in specific statements: " Collects information on client status using assessment skills of observation, interview, history taking, interpretation of laboratory data, mental health assessment, and physical assessment, including inspection, palpation, auscultation, and percussion" (College of Registered Nurses of Nova Scotia, (CRNNS), Entry-Level Competencies for Registered Nurses, item 31.)

Nursing programs prepare their students to meet national and provincial standards and once students graduate and pass the professional licensing exam the graduate is then considered to be a competent practitioner. This is a general expectation by the public and nursing regulatory bodies who state that registered nurses (RNs) are competent and will continue their professional development so as to maintain and enhance their competency. For example, the CRNNS has a continuing competence program for RNs and nurse practitioners (NPs) which, by its very title, suggests that both groups are competent once registered. Indeed, the CRNNS Continuing Competence Program is based on the philosophy that "RNs and NPs are competent and committed to lifelong learning" (CRNNS, Continuing Competence Program description, 2015). It is, therefore, realistic that nursing educators and researchers use the terms, competent or not competent, in measuring student outcomes during the clinical education process**.**

Discussion of SR Outcome Measurement Instruments

The SR sought to identify outcome measurement instruments used in high fidelity SBCE to measure undergraduate nursing students performance outcomes related to competence, CT, CJ, and CR. The SR findings revealed three main types of instruments: checklists, rubrics, and evaluation tools. Some of the tools were created especially for a specific SBCE study scenario (i.e. Simulation Thinking Rubric, Doolen, 2012) while others were designed to capture specific outcomes such as clinical judgment (Lasater Clinical Judgment Rubric, Lasater, 2005 & 2007) and were applicable for any SBCE scenario focused on that specific outcome. The discussion of the findings related to each type of instrument follows.

Lasater Clinical Judgment Rubric.

The Lasater Clinical Judgment Rubric (LCJR) was one of the measurement instruments that could be applied to different SBCE scenarios and that appeared most frequently in this SR (Ashcraft et al., 2013; Jensen, 2013; Lasater, 2005; Lasater, 2007; Meyer, 2012; Strickland, 2013). The LCJR tool created by Lasater (2005) is a rubric designed to capture and level students' CJ performance during a high fidelity SBCE single episode/event and has four subscales each with two or more identifying behaviour descriptor items: noticing (e.g. three items: focused observations, recognizing deviations, information seeking), interpreting, responding, and reflecting. The student's observed behaviour is graded on a 4 point Likert-type scale describing the level of student's CJ abilities: (1) beginning, (2) developing, (3) accomplished, and (4) exemplary thus allowing raters to identify CJ competency from simple to complex.

A feature of the LCJR is that it clearly presents an overall view of CJ development thus enabling students to grasp what CJ is and what it involves as the rubric language facilitates understanding of CJ expectations not only for students, but also for faculty, preceptors and LCJR raters (Lasater, 2007). This shared understanding of the meaning of CJ and how it is exhibited enhances joint understanding of the concept by all involved and could, therefore, translate to stronger interrater reliability. The LCJR's ability to apply to a variety of clinical contexts including long term care and community scenarios (as well adult acute care scenarios) is beneficial to educators in these practice areas where SBCE scenarios and measurement instruments may be more limited. Furthermore, the LCJR captures student performance more holistically than a checklist and appears to be more suited to higher level scenario complexity and performance with clinical reasoning because it examines task completion as well as clinical reasoning (Ashcraft et al. 2013).

CJ is a construct that is not easy to capture. Students struggle to understand CJ while educators struggle to explain and measure CJ. It is, therefore, significant that LCJR not only enables evaluators to capture this construct but also to determine differing levels of student CJ. It is also significant that SBCE provides a venue that offers appropriate, consistent, and time-appropriate experiences for all students and allows raters to measure student CJ. This is in contrast to clinical site practicum where instructors have intermittent (and often interrupted) time to observe and evaluate student CJ during student-patient events.

Other rubric instruments.

Five SR studies revealed rubrics developed especially for that study to measure competency outcomes which include: Doolen (2012), Gantt L.T. (2010), Haggard (2013), Nicholson (2010), and Swanson et al. (2011).

Simulation Thinking Rubric.

Doolen's (2012) self-developed instrument, the Simulation Thinking Rubric (STR) was designed to assess higher order thinking described as similar to clinical reasoning and judgment skills in nursing practice. The STR is a criterion-referenced measurement based on the Simulation Based on Learning Language model (SIMBaLL) learning theory that uses Piaget's four cognitive developmental stages. The rubric consists of four stages of thinking related to the simulation (sensorimotor, preoperational, concrete, and formal operations) and is scored from 1-7. While this instrument had a strong theoretical basis with clear operational terms they did not translate clearly into the instrument evidenced by the lack of construct validity and internal consistency. The STR was found to be unclear and difficult to score with overlapping and unclear empirical indicators in the cognitive developmental stage and with the theoretical basis of neuro-semantic learning language theory. Raters found there was an overabundance of qualifiers and mixed levels for scoring the student performance resulting in widely varying scores between raters. The STR was an unreliable instrument that did not indicate support for measuring higher order thinking.

Clark Simulation Evaluation Rubric.

In contrast, Gantt (2010) chose to measure student performance with the Clark Simulation Evaluation Rubric believing that rubrics capture more contextual and critical thinking components than do checklists. This rubric pairs Benner's five levels of nursing experience with

Bloom’s six cognitive domain categories such that each activity and level of performance on the rubric includes specific observable behaviours to assist the rater in scoring the student performance. This tool is simple and clearly laid out to evaluate student performance on the nursing domains of assessment, history taking, critical thinking, communication, patient teaching, and recognition of necessary diagnostic studies. Students are rated from 1-7 with a rating of 1 described as student “doesn’t see the picture” up to number 7 described as “anticipates the changing picture” (Gantt, 2010, p.102). Despite the overall simplicity of the tool the raters found the rubric language had variable definitions open to individual interpretation and that some performance areas seemed to overlap thus confusing the choice of score. Gantt noted that while the rubric is intended to grade groups of students at one time in a SBCE, her findings indicate that it was easier to score students individually with the tool.

It is interesting to note that a very detailed instrument such as Doolen’s STR shared the same rater interpretation issues as the Clark Simulation Evaluation Rubric which appears as a more concise and clear tool. This highlights a key issue with rating student performance with a measurement instrument, and that is ensuring that raters and faculty invest the time to become oriented to the tool, understand the wording, and behaviour descriptions and then discuss what constitutes successful and failing performances in that particular SBCE. This is essential to establish a level of interrater reliability (IRR).

Sweeney-Clark Simulation Performance Rubric.

Similarly, Haggard (2013) chose the Sweeney-Clark Simulation Performance Rubric (Sweeney & Clark, 2010). Haggard modified the tool by condensing the eight categories to focus more on safety, and align the categories to the Quality and Safety in Nursing Education (QSEN) competencies by including informatics which is not on the unmodified rubric. The categories are:

patient-centered care/assessment, teamwork, nursing interventions/evidence-based practice, communication, information gathering, critical thinking, safety, and informatics. Each category has a score range of one to five with one rated as beginner: “doesn’t see the picture”, to advanced beginner: “sees part of the picture”, to competent: “sees the picture”, then to proficient: “sees the big picture”, and at level five as expert: “anticipates the changing picture” (Haggard, 2013, pp.118-119).

This researcher compared the modified rubric with the Clark evaluation rubric (used in Gantt, 2010) and it was apparent that the Clark tool has brief general descriptors for each category / score level which would enable the rubric to be a broad spectrum SBCE measurement tool across many SBCE scenarios. However, there may be an explanation as to why the raters in the study had diverse scores. For instance, the category descriptors are so general that they are open to interpretation and leads one to question exactly what student performance behaviour is required for that category level. As well, some categories have wording and behaviours similar to another category thereby making it difficult to determine which category is the right one for scoring the student action. In contrast, the Sweeney-Clark tool descriptors for each category are detailed and specific with enough description to guide the rater clearly between performance levels and categories without losing the generalizability of the tool, thus making it applicable to various SBCE scenarios and to outcomes related to competence or critical thinking.

Student Performance Demonstration Rubric.

The Student Performance Demonstration Rubric was the tool of choice for Nicholson (2010), and Swanson et al. (2011) and is based on the American Heart Association guidelines for care of patients with myocardial infarction thus is limited to this cardiac scenario only. The rubric consists of 120 essential care items with a scale of zero to one, where zero indicates an inaccurate performance of essential care and one represents an accurate performance of essential care resulting in a total summed score for each student. The rubric is intended for use with a digital recording of the student performance and contains low-inference behaviours such that each item pertains to one separate and distinct behaviour so that little inference is required by the rater to determine scoring the performance (Nicholson, 2010).

The behavioural descriptors are very specific (e.g. heart rate assessed within five minutes, pain location assessed, and head of bed raised), and the rubric does include a comments column offering further detail for rater guidance which, in the opinion of this researcher, is necessary in the select sections in which they appear because the related care behaviour is open to inference in these areas (i.e. care element- diaphoresis; to be scored as yes or no; the accompanying comment is "states observation of diaphoresis" Nicholson, p. 150). Complete and accurate scoring of the rubric is dependent on hearing what the student is saying as some care elements can only be scored by what the student shares verbally by thinking aloud; it is also necessary to be able to observe all the student's actions, which is dependent on the camera view and which be problematic at times.

This researcher is experienced in marking recorded student performances and can attest to the need for careful attention in capturing the audio and video of such events. When the rater cannot hear what the student is saying or clearly see what the student is doing on video, then the rater has no choice but to give a zero score for that indicator behaviour. The rater cannot speculate that the student said the appropriate statement or carried out the action safely and accurately, thus camera and audio technicalities can cause students to lose points even when they are correctly performing the action. This can become a contentious issue when the student receives a low score for an evaluation item they know they performed but the rater could not mark. The rater or course professor is then placed in the untenable position of accepting both the rater's score and assessment of the student's performance or the student's claim as to what occurred which results in a devaluing of the evaluation process and reduction in reliability of the test outcomes. Nonetheless, the videotape of the student(s)' SBCE performance is particularly useful in the event of a student challenging the score they received (or did not receive) for a particular category. Lack of a video record of the performance can limit rater evaluation when needing to double check whether or not a competency was performed.

Clinical simulation evaluation tools.

Four of the SR studies included different versions of simulation evaluation tools: Aronson, Glynn, and Squires (2012), Mikasa, Cicero, and Adamson (2013), Patton (2013), and Radhakrishnan, Roche, and Cunningham (2007). The format of these tools were either a rubric or checklist but as they were titled simulation evaluation tools they are categorized separately in this study.

Heart Failure Simulation Competency Evaluation Tool (HFSCET) ©.

Aronson et al. (2012) developed a rating tool called the Heart Failure Simulation Competency Evaluation Tool (HFSCET) © to assess student nurse competency in responding to a deteriorating cardiac patient situation. Similar to the Student Performance Demonstration Rubric (Nicholson, 2010; Swanson et al., 2011) the HFSCET requires the students to "practice out loud" during the scenario telling the raters what they are thinking and doing, and what they are finding. This allows the raters to assess higher levels of cognitive functioning as well as psychomotor skills and to assess clinical reasoning in these and the other competency domains of patient safety, interventions, and documentation. Each domain is rated dichotomously and the tool includes brief descriptive statements that are specific and explicit (e.g. reads back and verifies physician order; dyspnea-must assess anterior and posterior pressure point) and appear to require little rater deliberation for scoring. A tool that is clear and quick to score is beneficial in a fast paced SBCE critical care scenario where the rater must watch, listen, and decide quickly on the score for an individual student or for a group of students. Raters lose valuable student performance viewing time when they are searching a measurement instrument for the right category or indicator to match the observed behaviour. This can cause the student to lose points on behaviours that occurred when the rater was not watching thus reducing validity of the test results.

The Seattle University Simulation Evaluation Tool.

Mikasa, Cicero, and Adamson (2013) created the Seattle University Simulation Evaluation Tool; a rubric style tool that integrated course objectives with the 1999 AACN baccalaureate competencies. This tool is designed to capture student indicators in the categories of assessment skills, critical thinking, patient care techniques, communication and collaboration within the student team, and professional behaviours. Each of the categories has a possible rating from 0-5 described as: 0 (below expectations) 1, 2, 3, 4 (no term provided) to 5 (exceeds expectations). There are three columns for each category such that each column has specific student behaviours required to achieve one of the two scores for that column (e.g. column one has a score of four or five, column two has a score of three or four, and column three has scores of one or zero).

Agreeing upon and describing professional behaviour indicators and outcomes for a SBCE (or for nursing in general) can be a challenging task for both educators and researchers, but this tool identifies and captures student actions that are appropriate for the SBCE and for a student professional level one. The professional behaviour category had specific indicators for the SBCE which were clearly defined for each level. For example: to receive a score of 4 or 5 for one of the professional behaviour indicators, the student demonstrates respect for client and team members; to receive a score of 2 or 3, the student shows respect for others inconsistently during a simulation event; and to receive a score of 1 or 0, the student did not demonstrate respect for client, peers or learning experience.

This researcher contends however, that selecting one of the two scores per column for a category would be problematic as there is only one set of behaviours listed, thus requiring the rater to make a judgment call in selecting the score. Individual rater decisions could lead to variations in rater scores even when observing the same student performance and thus negatively impact interrater reliability. The study raters did find the tool to be visual and efficient as they could circle the category behaviours quickly which is a critical factor in fast paced SBCE's.

A valuable feature of this tool (and the other quantitative tools included in this study) is the quantitative scoring system which is based on objective data to determine clinical grades rather than subjective rater comments ending with a pass or fail grade. It is possible that two raters can watch the same student performance in a simulation lab and interpret the level of competence differently unless the tool has clear, descriptive behavioural statements matched with a score. In the experience of this former nursing clinical laboratory instructor it was common for different raters to score the same performance differently due to several factors: (a) whether the rater is familiar with the student from other courses or laboratory sessions, (b) the rater's personal understanding of the testing objective and, especially, the tool categories, and (c) the degree to which the rater agrees with the required behaviours. Any one of these factors can lead the rater to stray from the tool's score and be more or less lenient in selecting a score.

Debriefing is a key component of SBCEs where the students and reviewer come together to discuss what happened in the SBCE, reflect on their own performance and that of their peers, as well as to determine how and if the SBCE student outcomes were met. The raters in this study noted that the tool was a useful guide during the debriefing conversation post-simulation to reinforce the scenario objectives and review the behaviours for each category. Use of the measurement instrument as a discussion guide provides the rater with clear discussion guidelines to follow and helps to ensure that all raters are covering the same general points with each group of students while also focusing on the actual events of that particular SBCE event. This method also helps students to integrate the learning from the event into the big picture of the SBCE learning objective(s).

The Creighton Simulation Evaluation Tool.

Patton (2013) selected the Creighton Simulation Evaluation Tool (C-SEI) to assess the categories of assessment, communication, critical thinking, and technical skills. Twenty-two behaviours were included in this tool and the score is dichotomous: zero for "does not demonstrate competence" or one for "demonstrates competence" (p. 194) and the rater can score a behaviour as "not applicable" if it is not included in the scenario. The final score is based on the percentage of competencies successfully demonstrated.

Inconsistent ratings occurred with this tool especially for the item "obtains pertinent objective data" (p. 195) which Patton suggests is due to lack of consensus on which data should have been collected by the students. This finding is related to the short time frame (one hour) allotted for rater training with the C-SEI, which was a major study limitation. Providing adequate preparation or training for SBCE raters using a measurement instrument is critical, yet the SR shows that this step is often overlooked or given minimal attention. Minimal training for SBCE raters has a direct negative impact on the reliability of the SBCE results as thorough training of evaluators is the most effective method of ensuring reliability with observational scales (Polit & Beck, 2008).

The Clinical Simulation Evaluation Tool (CSET).

Radhakrishnan, Roche, and Cunningham (2007) assessed nursing students on their performance in the categories of safety, basic assessment skills, prioritization, problem-focused assessment, interventions, delegation and communication measured with the Clinical Simulation Evaluation Tool (CSET). This tool is checklist style with check boxes for each indicator per category and includes a column where the specific behaviours are listed. The tool identifies the points for each specific behaviour. For example, under the safety category one indicator is called error-detect and interrupt. The SBCE specific behaviour items for this indicator are: wrong IV, boots off and O2 saturation off. The student receives a point for each item they identify and correct for a possible total of three points.

The format of this tool should help to increase interrater reliability because the tool behaviours are clear and specific and the points are already attached to each behaviour, therefore no judgment call is required on how many points to assign to the performance. This method of scoring also provides objectivity of the results as raters mark the behaviour as present or absent rather than a subjective evaluation of the correctness of the performance.

Checklist studies.

Four SR studies measured competency outcomes via checklist-type instruments. These studies are: Goodstone and Goodstone (2013), Kim and Shin (2013), Liaw et al. (2010), and Merriman, Stayt, and Ricketts (2014).

Medication Administration Safety Assessment Tool (MASAT).

Goodstone and Goodstone (2013) developed a performance-based competency measure of medication administration safety termed the Medication Administration Safety Assessment Tool (MASAT). The tool is an eight-item dichotomous checklist (possible student score of 0-8) measuring student behaviour adherence to the six rights of medication administration and safety. Due to the fact the tool items reflect medication rights any score below 8 represents a medication rights error meaning the student would fail and require remediation.

Each of the eight items relates to a medication safety right (e.g. right medication to right patient by asking the patient to state their name) and is written in brief, specific statements. The tool is designed for a single medication administration so a rater would need to complete a new tool for each medication given which could be cumbersome in a multi-medication SBCE. The method of establishing interrater reliability in this study provides both rater training and interrater accuracy checks on instrument use, and on scoring through the use of three seeded student performances (established by the researcher) and scored independently. These seed performances were videotaped behavioural samples with predetermined student errors with an established score given by the researcher/ expert. Rater agreement with the researcher's expert judgment introduces the validity of rater scores that is not possible with measures of inter-rater agreement (Johnson et al., 2009 in Goodstone & Goodstone, 2013) as comparison of individual ratings to expert ratings provides an assessment of the rater accuracy and indication of the tool's accuracy. Seeded examples are excellent rater training tools prior to testing and to assess rater drift on performance measures. The seeded samples then provide a baseline from which to monitor and provide corrective feedback to raters (Johnson et al., 2009 in Goodstone & Goodstone, 2013). Simulation plays an essential part in the production of the seeded samples and, therefore becomes a part of assessing its own ability to impact student outcomes as predicted by Gaba (2004).

The MASAT is unique in that it shows potential to measure translation of medication administration skills from the nursing simulation laboratory to patient care because it measures learning at Kirkpatrick's (1994 in Goodstone & Goodstone, 2013) evaluation level two (measure of learning) and three (measure of behaviour) and at translational science phase 1 (results achieved in educational laboratory) and at phase 2 (transfer to improved downstream patient care practices) (McGaghie, Draycott, Dunn, Lopez, & Stefanidis, 2011 in Goodstone & Goodstone, 2013). This is a significant finding because this level of evaluation has been lacking in SBCE research as most simulation evaluation instruments focus on low-level learner outcomes such as reaction and cognitive learning rather than the higher levels of participant's behaviours and patient outcomes (Adamson, Kardong-Edgren & Willhaus, 2013). Measuring SBCE student outcomes at the higher levels of evaluation will provide nursing researchers, educators, and stakeholders with an indication of the transferability of skills to the clinical setting; an important issue in preparing nursing students for practice.

Unnamed obstetrical SBCE checklist.

Kim and Shin (2013) developed an unnamed, 15 item checklist to evaluate students' competency skills and attitudes in various obstetrical nursing simulation scenarios under the categories of assessment, technical skills, patient teaching, prioritization, communication, patient privacy, and patient safety. With this tool, the rater is required to score the presence or absence of the expected action and then score the level of correctness for the action performed. The correctness scores range from zero (action was not performed) up to a score of three (accompanied by behaviour indicators). While this tool is unique in its design to score the required action and the ability to perform the action separately, this method could be problematic for raters unless they knew the tool behaviours well and did not have to scan the list each time to identify the right level for the behaviour observed.

Unnamed cardiac and respiratory SBCE checklists.

Liaw et al., (2010) had two sets of unnamed checklists to evaluate nursing students' cardiac and respiratory clinical crisis management performance. Each test checklist was designed for one of two test scenarios and included two subcategories: assessment and immediate actions, with a score range of one to three points (one point for no attempt, two points for an unsuccessful attempt, and three points for a successful attempt). The checklists are brief, specific to the SBCE situation, and list general behaviours on which to assess the student behaviour. For example, under the category immediate intervention one of the behaviours to score is "reassure patient" (p. 406) which does not specify certain expected behaviours but leaves the score open to the rater's interpretation and judgment of what patient reassurance behaviours are appropriate thus creating the likelihood of decreased interrater reliability. This tool is not developed as thoroughly as the other tools included in this SR.

Unnamed deteriorating patient condition checklist.

Merriman, Stayt, and Ricketts (2014) developed an unnamed checklist, which consisted of 24 objective performance criteria based on the Airway Breathing Circulation Disability Exposure (ABCDE) assessment tool. The tool indicates patient assessment areas, the expected order of priority of the assessment and the appropriate ensuing interventions. Although the checklist tool was used in the study there was insufficient data provided to enable a tool evaluation by this researcher.

SBCE Tool Summary

The SR outcome-based measurement instruments all provide a quantitative score based on objective data which is a benefit to both the students and raters as grading clinical performance tends to be a highly subjective process (Mikasa, Cicero, & Adamson, 2013).

The study instruments shared strengths in instrument design such as measuring outcomes in a manner that enables their use in a variety of SBCE scenarios (dependent on which competence outcome the educator wishes to evaluate), capturing higher level thinking and professional behaviour.

The use of low inference language in describing the student performance indicators is a strength in that it enables raters to accurately and quickly locate the student performance behaviour on the tool as the SBCE is running. These tools were notable for this feature: the LCJR, Modified Sweeney-Clark Rubric, Clinical Simulation Evaluation Tool, and the MASAT.

While each of the tools included in this SR review have both strengths and weaknesses there are tools that: are strong in capturing and measuring student competence in more general and in comprehensive scenarios (not health condition specific such as cardiac events only), have low inference language, and provide a unified score for presence of the action and level of the student's ability to perform the action. Findings from this SR point identify the Lasater Clinical Judgment Rubric and the Seattle Evaluation Tool as strong in these areas.

The LCJR would be most useful in SBCE scenarios with higher-level BN students (year three or four) who have intermediate levels of CJ ability, so the evaluator is able to assess their intervention skills and clinical judgment as they respond to a complex care situation. For example, scenarios such as a deteriorating patient situation post-surgery, anaphylactic shock, chronic renal failure requiring astute assessment, interventions and clinical judgment. Nursing schools would find this tool useful in assessing student practice and evaluation for high-acuity low-occurrence events (such as cardiac and respiratory crisis) thus ensuring all students have experience in such a situation prior to graduation.

The Seattle Evaluation Tool could be applied to any type and level of SBCE seeking to capture CR, CJ, and professionalism as it captures these features more strongly than basic intervention skills. For instance, this tool would be useful in a mental health, family, or community scenario where the focus is more on the student's ability to assess, communicate effectively and professionally, and jointly develop a plan of action rather than capture psychomotor skills (the tool can capture psychomotor skills but it is not its strength according to the findings of this SR). While the tool may need modification for certain scenarios it provides nursing programs with an instrument for scenarios in mental health, family, and community clinical practice areas which currently are not as well developed as medical-surgical acute care scenarios.

The health-condition specific tool that is best for assessing cardiac SBCE events is the HFSCET as it captures the actions precisely, is quick and clear to score, and allows the rater to capture the students' clinical reasoning as they are a required to "practice out loud" sharing their thoughts with the rater. This would be an excellent tool to provide practice and or assessment for final year students in this critical area.

The MASAT shows great promise for SBCEs involving a focus on medication administration skills as it measures clinical performance skills that translate to patient care, however the tool would need modification to allow the rater to assess the students' performance in administering more than one medication at a time. This tool could be used in conjunction with the Seattle Evaluation Tool during a second year student SBCE in which the student(s) must assess and give pain medication(s) prior to performing wound care.

In the future, we need a tool that is adaptable/suitable to any level of student, to various clinical settings (both acute care and community), has low inference indicators, and measures competence and CJ, CR, and CT as they are all part of the competence process. The literature review and SR have shown that nursing researchers have developed many different evaluation tools to measure SBCE outcomes, and while some of them have been tested in more than one study with a new population, many have not. The evidence also shows that it is time to stop creating new tools and focus on retesting those tools already in use. This researcher is in agreement with this statement because the SR findings indicate that the wealth of tools already developed could be applied to a broad spectrum of SBCE scenarios seeking to measure student outcomes of CT, CJ, CR, and competence.

SBCE Instruments Reliability and Validity

Another factor in choosing a tool is evidence of the tool's validity and reliability from its use in a previous study or SBCE. Tools with established validity and reliability are especially important when the measurement instruments are used in SBCE for the purpose of high stakes testing in order to ensure that each student is being tested on the same construct in a reliable manner such that if two students have similar performances their scores should also be similar. This researcher contends that all students involved in a SBCE experience or testing event expect that the measurement instrument is fair (measures all students to the same standard), that all raters are looking for and scoring the same behaviour and performance the same way, and that the rater's personal judgment does not enter the scoring decision. The only method of ensuring that this is the case in each SBCE is to continually test and retest the measurement instrument and its psychometric properties with different populations and scenarios. The SR study findings on measurement instrument validity and reliability are presented in Chapter four; this summary will focus on key findings from these results.

Validity of SBCE instruments.

Content validity was the most frequently reported measure of validity for the studies (13/19 studies) and half of them used the single method of literature review while the other studies also included a second method of validity testing: a review of the tool by a content or instrument development expert or panel of experts. There is a lack of content validity evidence reported by researchers of SBCE instrument studies based on content validity index statistics (CVI), which is concerning for two reasons: first, because it provides a precise measure of content validity regarding the relevance and appropriateness of the tool items to measure the construct, and second because nurse researchers have been instrumental in developing an approach that involves the calculation of a CVI by an expert panel (Polit & Beck, 2008).

As the current best measure of content validity is the review and evaluation of the instrument by a panel of substantive content experts (SCE), it is recommended that CVI statistics become the next step in standard content validity testing as it too involves the use of an expert panel, which many researchers are already using. One study did report the use of factor analysis to help establish content validity which is also useful in identifying the interrelationships among instrument items and confirms that they fit together as unified concepts (Polit & Beck, 2008). Ideally, a study would include two to three of these measures, so it is time to include CVI and/or factor analysis in routine testing for instrument validity in SBCE instrument research. These findings align with other researchers' findings of SBCE instruments (Kardong-Edgren et al., 2010).

Construct validity is considered the most important of the validity properties indicating that the tool is measuring the hypothetical basis for the variable, however this is often abstract and hard to measure (Houser, 2012) but fundamental to a strong evidence base (Frasure, 2008). Construct validity is particularly important in situations where tests are evaluating attributes that are not easily or objectively measured (i.e. CJ), and one study (Haggard, 2013) reported construct validity measures which validates that the instrument is actually measuring the instrument construct adequately. This finding is not surprising as Adamson et al. (2013) note that it is very difficult to establish validity of performance-based tools as they are frequently subject to the perceptions, knowledge, experience and training of the evaluators. This is an area requiring researcher attention in future SBCE research.

Reliability of SBCE instruments.

The SR found that eight of the 19 studies did not report any reliability measures for the use of the instrument in their study. While this is an expected finding it points to a serious issue in SCBE instrument research because reliability indicates the consistency with which a tool measures the chosen construct and the accuracy of the measure. Without reliability testing the researcher cannot be sure that the instrument items are all measuring the same trait (internal consistency); that the raters or observers agree on the scoring of the instrument (interrater equivalence); and that the instrument is stable among participants, and over time (test-retest reliability). All of these are critical assessments of an instrument's reliability and should be tested each time a researcher uses the tool because the tool's reliability is not inherent; its reliability depends on how the tool is applied to a certain sample and the circumstances (of the SBCE), so new estimates of reliability are recommended.

Some SR researchers quoted tool reliability findings from previous research studies, however this is not sufficient. It is unclear why researchers are not conducting reliability measures on the tool, however it is probable that it is due to lack of financial, human, and or time resources. Lack of reliability measures impacts the tool greatly because an instrument that is unreliable cannot be valid either, so the tool cannot be considered an accurate or useful measure (Polit & Beck, 2008). For instance, Lasater 2005 and 2007 did not report any reliability findings, but did report content validity findings; which decreases the strength of the study findings as both reliability and validity are required to trust the outcome of a study (Houser, 2012; Polit & Beck in Frasure, 2008).

Test - retest stability reliability measures were reported for two of the 19 studies which is a good step towards improved psychometric measurement reporting as weak stability reliability indicates that the tool may not be reliable in providing similar results in repeated testing situations.

When the SR researchers did report reliability measures, internal consistency statistics were the most common, which is not surprising because it is the most widely used reliability approach among nurse researchers (Polit & Beck, 2008). This measure is particularly important in SBCE because a strong internal consistency (or test retest reliability) is essential when assessing students' performance during simulation scenarios (Shelestak & Voshall, 2014) and in high stakes testing where the alpha coefficient should be 0.9 or greater. A strong internal consistency ensures that each time the tool is administered it is measuring the same construct for each student. SR studies with reported Cronbach values indicated that only five of the studies which reported an alpha value had an alpha coefficient at the recommended level for SBCE testing. Because these tools determine student's clinical competency it is important that they are reliable and valid (Decker et al., 2008).

Instrument and interrater reliability (IRR) is also a critical measure for SBCE instruments as a high level of agreement on instrument scoring indicates an assumption that measurement errors have been minimized (Polit & Beck, 2008). It was surprising to find that few SR studies (six) included IRR evidence, as accurate scoring is vital to the integrity of the SBCE, and student outcomes, and because there are many methods available by which to establish IRR. However, this is an increase in the reporting of IRR from previous nursing research instrument studies which showed there was no reporting of IRR (Kardong-Edgren et al., 2010), so it appears that SBCE instrument researchers are beginning to place an emphasis on this in current studies. It is worth noting that IRR can be difficult to establish in SBCE due to the changing scenario in which the student behaviours and actions may change rapidly as well making it difficult for a rater to capture and evaluate (Kardong-Edgren et al., 2010). Based on the fact that only five of 19 studies reported a Cronbach alpha at or above 0.9 (which is the recommended statistic necessary to ensure a strong internal consistency for high stakes testing) and the finding that few SR studies (six) included IRR evidence to indicate accurate scoring (vital to the integrity of the SBCE and student outcomes) it appears that nursing educators cannot yet use SBCE to evaluate readiness to practice via high stakes simulation testing.

Based on the reliability and validity findings of this SR the following instruments have at least one or more measures of validity and reliability on which to consider them as valid and reliable tools to retest and use in an SBCE; (a) the HFSCET(Aronson, Glynn, & Squires 2012), (b) the Modified LCJR (Ashcraft et al., 2013), (c) the MASAT (Goodstone & Goodstone, 2013), (d) Sweeney-Clark Simulation Performance Rubric: Haggard Modification (Haggard, 2013), (e) The Seattle University Simulation Evaluation Tool (Mikasa, Cicero & Adamson, 2013), (f) Student Performance Demonstration Rubric (Nicholson, 2010), (g) The Creighton Simulation

Evaluation Instrument (C-SEI) (Patton, 2013), (h) Student Performance Demonstration Rubric (Swanson et al., 2011). The SR tools that have both reliability and validity findings that meet standard statistical requirements are the Medication Administration Safety Assessment Tool (Goodstone & Goodstone, 2013), The Seattle University Simulation Evaluation Tool (Mikasa, Cicero, & Adamson, 2013), and The Student Performance Rubric (Nicholson, 2010).

SCBE measurement instrument research is a fairly new area of research in nursing and one that is fairly complex in comparison to simulation research in related fields of health education such as medicine, respiratory therapy, and physiotherapy. These fields of practice focus primarily on the dynamics of physical assessment and physical skills which tend to be observed and measured more easily, whereas nursing simulation research has a broader focus on physical skills as well as a holistic client assessment and care including families and communities. Social science researchers in such fields as psychology have expertise in tool development and stress psychometric assessment in related research, whereas this has not been the case in nursing research where SBCE researchers have tended to adapt tools from these sciences rather than develop one.

COSMIN Quality Assessment of Studies

A strong study design is important to ensure that the study results capture the intended variable and the results are representing the evidence in an objective and truthful manner on which other scholars and researchers can base decisions. Various quality assessment tools exist to assist researchers in assessing the quality of primary research studies and are selected by taking into consideration the type of study design (e.g. experimental, qualitative, quantitative), and other essential quality domains. Despite the large number of quality assessment tools available, a lack of consensus exists regarding which tools are most appropriate for nurse researchers largely because there is no “gold standard” for determining scientific rigour and validity of primary studies (Polit & Beck, 2008). Due to the lack of a gold standard it is therefore difficult to validate the assessment instruments as well.

While this researcher found various quality assessment tools within nursing research studies, many of them were not aimed at assessing the quality of the study in terms of the study design and the psychometric properties of the SBCE outcome measurement instrument used in the study. This is not unusual as quality assessment criteria vary widely from one instrument to another (Polit & Beck, 2008) thus providing different ratings depending on which quality tool is used. The COSMIN quality assessment instrument (Terwee et al., 2012) was selected to assess the study quality of the SR studies because the COSMIN assessment checklist was designed to evaluate the methodological quality of studies on measurement properties of health related, patient-reported outcome instruments and is also recommended for studies measuring performance-based outcomes. This researcher found the COSMIN study quality criteria also appropriate for assessing the quality of student performance outcome measurement instruments.

COSMIN ratings for the SR studies tended to be low and appear to be in contradiction with the statistical findings for validity and reliability findings for many of the individual studies. This is due to the fact that the COSMIN criteria are inclusive of many factors. For example study criteria for reliability include:

- The percentage of reported number of missing items.
- The presence of a description noting how missing items were handled.
- Sample size (anything below 30 provides a rating of poor).
- Whether or not two measurements were available, and if so, were they administered independently with an appropriate time interval stated.

- Whether or not the patients (students) were stable on the construct being measured in the interim between administered tests.
- Whether or not the test conditions were similar for both measurements.
- Whether or not an intraclass correlation coefficient was calculated for continuous scores.
- Whether or not kappa was calculated for dichotomous/ordinal/nominal scores.
- Whether or not a weighted kappa was calculated for ordinal scores.
- Whether or not a weighting scheme was provided for ordinal scores.

Any one rating of poor for any of these criteria results in a rating of poor for the study reliability as a whole. So, even if the study meets statistical reliability standards if it does not meet COSMIN standards the study receives a rating of poor with this tool.

The SR findings show that all 19 studies received a rating of poor for internal consistency reliability. The reasons for the poor rating for internal consistency was due to either one or a combination of the following: (a) a small sample size (< 30), (b) the lack of factor analysis or reference to another study with same, (c) the lack of Cronbach's statistic, (d) the lack of an internal consistency statistic for each subscale of the tool, and (e) lack of a goodness of fit statistic at a global level.

Sample size challenges face SBCE nursing researchers that include student recruitment availability, very small or very large class sizes, ethics considerations regarding the use of students as research participants, student attrition and program requirements that may cause diffusion of treatment issues. For instance, stability of students' competence may be affected during SBCE research when some of the group have their required clinical practicum at the same time, thus impacting their competence.

There is lack of factor analysis in SBCE studies and a possible reason is that factor analysis requires a larger sample size than is needed for Cronbach's statistical analysis and it may be difficult for the nursing SBCE researchers to increase sample size to allow for this. Cronbach's alpha is the best means of assessing measurement error in psychosocial instruments (Polit & Beck, 2008) and is also more feasible, especially for nursing researchers because it requires only one test administration thus is resource friendly in terms of time and human resources.

While it is common (and often unavoidable) for nursing research studies to have smaller sample sizes, the COSMIN quality assessment findings reinforce the need for larger sample sizes. The findings also speak to the need for more thorough statistical testing and reporting in order to establish strong internal consistency in studies that test measurement instruments.

Similarly, the SR studies received a rating of poor quality for reliability which was due to either one or a combination of the following: (a) no description of the number of missing items and/or how missing items were handled, (b) small sample size, (c) lack of an intraclass correlation coefficient (ICC) or Pearson or Spearman correlations, (d) lack of Kappa statistic for ordinal, dichotomous, or nominal scores, (e) lack of two independent measurements, (f) lack of stability of the subjects between repeat tests, (g) and test conditions not similar.

The issue of reporting the number of missing scored items and how this was handled is a recurring item in the COSMIN assessment as it is designed for health related patient surveys. Raters of student performance measurement tools tend to score every item, therefore missing item scores are not an issue. If one item was not scored it likely meant that the student did not perform the required action and received a score of zero. Nonetheless, to keep the integrity of the COSMIN assessment boxes it was necessary to include this item in assessing the studies.

COSMIN standards for reliability state that in order to evaluate reliability of the instrument it should have been administered twice in independent tests and specifies that the rater and subject must both be unaware of the scores on the first test. This was not the case in any of the studies, and while it would be possible to keep the first test score from the student, it may be considered unethical to do so, as students are generally anxious to know their score and understand where they went wrong so that they can then get remedial assistance or study further in the area of weakness. Students are especially anxious to receive their scores and performance comments when they are facing another test requiring similar skills and knowledge, as would be the case in the repeat testing scenario required by the COSMIN criteria.

Running a second test with a new set of raters would be difficult for nursing researchers. Securing adequate numbers of faculty raters to score one student performance test or SBCE scenario is challenging, let alone attempting to secure a second set of raters in order to have two different sets of raters (as required by the COSMIN criteria). It is the experience of this researcher that securing and ensuring the presence of faculty raters for student performance tests is an onerous task due to the busy schedules of teaching faculty and simulation lab faculty and due to the limited scheduling availability of students and laboratory times. It would be possible for the researcher to secure one group of raters for both tests, but it would be a challenging task requiring very careful scheduling of raters and students to ensure that the rater did not test the same student as in test one. For the above reasons, the COSMIN criteria for this aspect of reliability are currently unrealistic for SBCE nursing research studies. As previously discussed, reporting the use of statistical measures is minimal in SBCE research studies and this also impacted their COSMIN assessment ratings.

The SR studies also received a rating of poor for the measurement error criteria again due to sample size, and no percentage of missing items given, no repeat test administration, and due to a lack of stability of the participants. Nursing SBCE research studies could improve this rating by striving to include methods that would meet these criteria at a beginning level. For instance, instead of running two administrations of the instrument with two separate groups of testers, strive for two instrument administrations with the same testers and a reasonable interval between tests (i.e. beginning of term and end of term).

COSMIN content validity ratings was the strongest area for the SR studies and is not surprising as this is one area that most researchers reported on. Theoretical foundation of the instrument was lacking in several studies which caused them to be rated as fair under COSMIN standards.

Instrument generalizability rated from poor to good for the study instruments; the leading reason for poor ratings was the lack of reporting on the participants' mean age and gender, and diversity demographics. This would not be a difficult item for researchers to include and would clarify the ability of the tool to be used in other studies and populations. Once again the sampling method of convenience sampling caused a low rating.

If selecting a tool based on the instrument's COSMIN quality assessment, the SR evidence supports the Heart Failure Simulation Competency Evaluation Tool (Aronosn, Glynn & Squires, 2012), The Lasater Clinical Judgement Rubric (Strickland, 2013) and the Student Performance Rubric (Swanson et al., 2011) as each one has a rating of fair for measurement error and excellent for content validity. All three tools have a rating of poor for internal consistency and reliability.

Overall, the COSMIN tool was useful in guiding quality assessment of the studies because it was directly related to measurement instrument studies and focused on study design areas necessary for psychometrically sound instrumentation. While the tool was not applicable to nursing research per se, the COSMIN checklist could be adapted for SBCE instrument research studies. For instance, removing the missing items criteria and adjusting the sample size criteria to one that is more feasible for SBCE researchers but does not compromise study quality would enhance the usefulness and applicability of the checklist for use in future nursing studies.

Learning Domains and SBCE

The SR results showed that none of the 19 SR studies demonstrated specific performance parameters by which to identify the conative domain via the observation measurement instrument. It is not unusual that an observation measurement instrument would fail to capture the conative and affective domains because both domains involve the characteristics of self-confidence, self-regulation, motivation, volition, personal and professional values, and self-directed learning strategies, and are best captured by self- report tools (Huitt & Cain, 2005). Thus this researcher did not attempt to judge if these domains were met in the studies as the only available evidence were statement descriptors in the tool or statements in the discussion of findings neither of which were considered objective indications of the presence of the domain.

All 19 studies captured the cognitive and psychomotor learning domains which is understandable as most SBCEs are designed to capture these areas. Huitt and Cain (2005) identified that the four domains of learning are interconnected, so it is likely that some of the tools do capture the affective domain (and the conative domain), but neither domains were identifiable in this study due to the SR focus on strictly observable outcome behaviours. One exception that could have been made was the study where the “practice out loud” method

(Aronson, Glynn, and Squires, 2012) was used enabling the observer to determine the thinking behind the students' actions.

Final Note on Selecting Measurement Instruments

Many factors come into consideration when selecting a measurement instrument for use in SBCE practice and research: the ability of the tool to capture the identified SBCE concept and outcome; the instrument's previous reliability and validity findings; the methodological quality of the study testing the instrument; the tool's format and ease of use; and the ability to apply the tool to the desired population and scenario.

Study Strengths and Limitations

Study strengths.

The SR methodology itself is a strength of this study because a systematic review that uses explicit, systematic methods that are selected in order to minimize bias, can reflect all relevant, scientifically sound research thus providing more reliable findings from which conclusions can be drawn and decisions made (Higgins & Green, 2011).

This SR followed the SR standards of Preferred Reporting items for Systematic reviews and Meta-Analyses criteria (PRISMA) checklist and guide to ensure that the SR process and results are transparent and complete (Moher, Liberati, Tetzlaff, & Altman, 2009). Following PRISMA guidelines was a strength of the study because it provided guidance to the researcher and reinforced the need to: develop a well-focused and feasible question, create and follow an explicit protocol and methods for evaluating retrieved studies, and to create and employ the SR study selection sheet and data extraction sheet in order to provide a transparent description of search and selection methods that are reproducible by another researcher. The broad SR search was a strength in itself as it encompassed eight subject databases, two subject specific sources and two general search engines and had an open start time frame allowing inclusivity of earlier research up to November 2014.

Another strength of the SR was that two reviewers considered studies for all stages of the search, for study selection, for data extraction, and for the quality assessment of the studies. The reviewers showed a strong rater agreement on both study selection and quality assessment of studies.

The methodological diversity of studies in this SR is also a strength as there is agreement that a SR can, and should, use a range of different research designs. By doing so it allowed this researcher to include quantitative studies that were not pure experimental research but offered important evidence on instruments that would otherwise have been excluded. The process of including the studies of lesser quality at the quality appraisal stage allowed the researcher to consider threats to validity during the analysis and interpretation phases of the SR.

The choice of the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) checklist to assess study quality was a strength as its focus is on study quality as it pertains to the validity and reliability of the measurement instrument in the study. The COSMIN checklist is applicable to a variety of study designs and the assessment categories provided specific criteria and guidelines for use that allowed the reviewer to consider all facets of a study's quality and instrument validity and reliability.

Study limitations.

One limitation of the study was the absence of non-English studies (due to study selection criteria) which would have allowed for greater generalizability of SR results. A second limitation was that during the study selection process, 10-15 studies had no accompanying abstracts so the second reviewer excluded them due to lack of information. Once the researcher provided the abstract, the second reviewer independently reviewed the study abstracts for inclusion or exclusion and agreement was reached on inclusion or exclusion of each study. This step was outside the usual study review process and could be seen as inducing bias.

Another limitation was the lack of Canadian studies in the review. This was primarily due to the fact that the Canadian SBCE studies captured by the search were qualitative and therefore, did not meet the study criteria. There was one Canadian study that was quantitative but it was not retrievable. Excluding qualitative studies limits the SR findings by potentially missing information from other researchers' and undergraduate nursing students' experiences of SBCE evaluation methods and tools and in particular, excludes the Canadian experience with SBCE student evaluation.

The introduction of a new second reviewer occurred after the studies were selected and before assessment of study quality began. This occurred because the first reviewer was an undergraduate student (with a previous degree) and the study assessment required full-time work for a period of time that he could not commit to. Once a new reviewer was secured rater agreement was established by having her review the first five studies on the study list to review for SR inclusion or exclusion. The change of reviewers could introduce a new source of bias as this reviewer was not part of the original selection process.

Lastly, there were six near miss studies that could have been included in this SR as they met all the inclusion criteria but one: study population. The population for the SR was undergraduate nursing students in a baccalaureate or associate degree program, but the population of the near miss studies was faculty raters. Six studies that presented new information on instrument validity and reliability had to be excluded for this reason. This factor limits the SR's ability to represent all relevant SBCE measurement instrument research. Of course, this is balanced by the fact that in a systematic review there is always the threat of possible valid studies not being included due to the specificity of the search and due to the databases themselves and their limited inclusivity of studies.

Significance of the Research

To the best of this researcher's knowledge, other studies have not attempted to identify specific performance behaviours for competence, CT, CR, and CJ from the literature and then apply them to current SBCE research on measurement instrument outcomes in order to assist in identifying the tools that capture each performance behaviour best. This method allowed the researcher to quantify the behaviours captured for each outcome and then provide a comparison of the type and number of outcomes captured by each tool. This method could be helpful to future researchers and SBCE educators when selecting a tool that best captures the desired outcomes of their SBCE.

During the literature review of the definitions and use of the terms competence, CT, CR, and CJ, the researcher learned that this is an area that lacks consensus and has an impact on SBCE measurement instrument research. In exploring the literature on this topic the researcher found evidence of the interrelated nature of the four processes of competence, CT, CR, and CJ involved in competent nursing practice. This evidence became the basis for a conceptual model.to represent and explain the interrelatedness of the concepts to competent nursing practice.

As well, this research provided evidence of five common indicators used in SBCE instrument research that can be considered basic indicator components in measuring student performance outcomes of competence. This could be a step toward consensus on a definition of nursing competence.

This study introduced a quality assessment tool not previously used in SBCE research but is one that shows promise for future SBCE measurement instrument research and indicates the importance of reporting instrument validity and reliability measures.

SR Implications for Nursing Research, Practice, and Education

Based on the results of this SR, focused discussions are needed by nursing stakeholders (researchers, educators, students, practitioners, regulators, and consumers) to reach consensus on what nursing competence is, how it is defined, and how the processes of CT, CR, CJ, and competence impact the outcome of competent nursing practice. This would also include discussion related to the level of performance that indicates competence for nursing students and graduate nurses.

Further to this, SBCE research could investigate the use of more than one tool to measure student performance outcomes based on the fact that one tool may not be able to capture several different competencies that occur within different domains of learning.

This SR echoes the recommendations of previous researchers that new SBCE measurement instruments should not be developed at this time; instead the focus needs to be on retesting the existing tools with larger samples and in new populations. The SR findings indicate that existing tools already developed can be applied to a broad spectrum of SBCE scenarios seeking to measure student outcomes of CT, CJ, CR, and competence. As well, these tools could be adapted cater to any level of student, to various clinical settings (both acute care and community), and includes low inference indicators, all of which are desirable features in a tool that can be used widely in SBCE research and education.

Based on the SR finding, the Seattle University Simulation Evaluation Tool© is the tool recommended for SBCE undergraduate nursing student outcome evaluation for various reasons:

the tool was included with two other instruments in a multi-site collaboration assessing reliability of simulation evaluation tools (Adamson & Kardong-Edgren, 2012), the Seattle tool has a reported internal consistency reliability of Cronbach’s alpha .97 (Mikasa et al., 2013), and is one of the SR study tools that met reliability and validity standards. As well, the Seattle tool is designed to capture student indicators in the categories of assessment skills, critical thinking, patient care techniques, and communication and collaboration within the student team, and professional behaviours. Additionally, the SR results show that the tool captured the concepts of CR (5/6 outcomes) and CT (4/5 outcomes), competence (3/5 outcomes), and CJ concept (3/8 outcomes). It is also a flexible tool for use with various scenarios and has added value by identifying and capturing student actions that are appropriate to measure professional behavior outcomes at the student level via specific level indicators for the SBCE. The Seattle University Simulation Evaluation Tool© was found to be a useful guide during the debriefing conversation post-simulation to reinforce the scenario objectives, to review the behaviours for each category, and to provide the rater with clear discussion guidelines to follow which helps to ensure that all raters are covering the same general points.

The SR also showed that researchers would reap the benefits of stronger interrater reliability by providing increased time and discussion in orienting instrument raters, perhaps with the use of seeded samples of student performance.

SBCE educators and researchers could assist in the science of translational research by moving SBCE research and practice outcomes from the clinical simulation laboratory to the care of the patient thus affecting health outcomes. This can be done by choosing or designing scenarios, research, and instruments that allow the demonstration and evaluation of student performance at the higher levels of evaluation: translational phase two and three. Activities at these levels assess if what the students learned in the SBCE will carry over into a patient care setting and if what was demonstrated in the SBCE will carry over to the patient care setting resulting in improved health outcomes.

Conclusion

The evidence shows that SBCE is used frequently for evaluating competency outcomes, often in patient scenarios that include rapidly deteriorating patient conditions requiring astute judgment and quick action by the student nurse. Nursing students rarely have the opportunity to participate in these high acuity situations during on-site clinical practice due to patient safety reasons, thus clinical instructors do not often have the opportunity to evaluate student CJ under such circumstances. In fact, with the increasing number of students and the decreasing availability of clinical sites accepting students, it is necessary to provide other clinical practice and evaluation opportunities for nursing students.

As a clinical coordinator requesting close to 2000 student placements per calendar year it is obvious that clinical sites cannot accommodate student numbers, nor provide the necessary clinical practice experiences to provide all students with the patient care experiences necessary for clinical course objectives. High-fidelity SBCE is now the practice method of choice for our second year baccalaureate students at Dalhousie University School of Nursing (Halifax, Nova Scotia) during their second term to consolidate their year two competencies prior to initial acute care experiences later in the year. This is a step in focusing on what ultimately matters in simulation: producing safe, competent practitioners who positively impact patient outcomes during clinical practice.

References

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

Adamson, K., Kardong-Edgren, S., & Willhaus, J. (2013). An updated review of published simulation evaluation instruments. *Clinical Simulation in Nursing*, 9(9), e393. doi:10.1016/j.ecns.2012.09.004

Alinier, G., Hunt, W. B., & Gordon, R. (2004). Determining the value of simulation in nurse education: Study design and initial results. *Nurse Education in Practice, 4*(3), 200-207. Doi:10.1016/S1471-5953(03)00066-0

Alinier, G., Hunt, B., Gordon, R., & Harwood, C. (2006). Effectiveness of intermediate-fidelity simulation training technology in undergraduate nursing education. *Journal of Advanced Nursing, 54*(3), 359-369. Retrieved from http://ezproxy.library.dal.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true

Aronson, B., Glynn, B., & Squires, T. (2013). Effectiveness of a role-modeling intervention on student nurse simulation competency. *Clinical Simulation in Nursing, 9*(4), e121-6. doi:10.1016/j.ecns.2011.11.005

Ashcraft A.S., Opton L., Ruth, B. A., Caballero S., Veesart A., & Weaver C. (2013). Simulation evaluation using a modified lasater clinical judgment rubric. *Nursing Education Perspectives, 34*(2), 122-126. Retrieved from http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L368948884; http://dx.doi.org/10.5480/1536-5026-34.2.122; http://sfxhosted.exlibrisgroup.com/dal?sid=EMBASE&issn=15365026&id=doi:10.5480%2F1536-5026-34.2.122&atitle=Simulation+evaluation+using+a+modified+lasater+clinical+judgment+rubric&stitle=Nurs.+Educ.+Persp.&title=Nursing+Education+Perspectives&volume=34&issue=2&spage=122&epage=126&aulast=Ashcraft&aufirst=Alyce+S.&auinit=A.S.&aufull=Ashcraft+A.S.&coden=&isbn=&pages=122-126&date=2013&auinit1=A&auinitm=S.

Bambini, D., Washburn, J., & Perkins, R. (2009). Outcomes of clinical simulation for novice nursing students: Communication, confidence, clinical judgment. *Nursing Education Perspectives, 30*(2), 79-82. Retrieved from http://ezproxy.library.dal.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=c8h&AN=2010258640&site=ehost-live

Banning, M. (2008). Clinical reasoning and its application to nursing: Concepts and research studies. *Nurse Education in Practice*, 8,177-183.

Barnsley, L., Lyon, P. M., Ralston, S. J., Hibbert, E., Cunningham, I., Gordon, F. C., & Field, M. J. (2004). Original article clinical skills in junior medical officers: A comparison of self-reported confidence and observed competence. *Medical Education,* 38(4), 358-367. doi:10.1046/j.1365-292

Baxter, P. & Norman, G. (2011). Self-assessment or self-deception? A lack of association between nursing students' self-assessment and performance. *Journal of Advanced Nursing,* 67(11), 2406-2413. doi:10.1111/j.1365-2648.2011.05658.x

Benner, P. (1982). From Novice to Expert. *American Journal of Nursing 82*(3), 402-407.

Bettany-Saltikov, J. (2012). *How to do a systematic review in nursing. A step-by-step guide.* Berkshire, England: Open University Press.

Black, D., Allen, L., Redfern. (2008). Competencies in the context of entry-level registered nurse practice: a collaborative project in Canada. *International Nursing Review, 55*, 171–178.

Bland, A. J., Topping, A., & Wood, B. (2011). A concept analysis of simulation as a learning strategy in the education of undergraduate nursing students. *Nurse Education Today, 31*(7), 664-670. doi:http://dx.doi.org/10.1016/j.nedt.2010.10.013

Blum, C. A., Borglund, S., & Parcells, D. (2010). High-fidelity nursing simulation: Impact on student self-confidence and clinical competence. *International Journal of Nursing Education Scholarship, 7*(1), 14p. doi:10.2202/1548-923X.2035

Brewer, E., P. (2011). Successful techniques for using human patient simulation in nursing education. *Journal of Nursing Scholarship, 43*(3), 311-317. doi:10.1111/j.1547-5069.2011.01405.x 3.2004.01773.x

Brunt, B. (2005a). Models, Measurement, and Strategies in developing critical –thinking skills. The Journal of Continuing Education in \nursing, 36(6)255-262.

Brunt, B. A. (2005b). Critical thinking in nursing: An integrated review. *Journal of Continuing Education in Nursing, 36*(2), 60-67. Retrieved from http://ezproxy.library.dal.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=c8h&AN=2005079207&site=ehost-live

Buykx, P., Kinsman, L., Cooper, S., McConnell-Henry, T., Cant, R., Endacott, R., & Scholes, J. (2011). FIRST^{2}ACT: Educating nurses to identify patient deterioration - A theory-based model for best practice simulation education. *Nurse Education Today, 31*(7), 687-693. doi:10.1016/j.nedt.2011.03.006

Canadian Patient Safety Institute, (2008). Patient Simulation Needs Assessment. May 19, 2008.

Cant, R. P., & Cooper, S. J. (2010). Simulation-based learning in nurse education: Systematic review. *Journal of Advanced Nursing, 66*(1), 3-15. doi:http://dx.doi.org/10.1111/j.1365-2648.2009.05240.x

Carlson, E. (2011). Ethical considerations surrounding simulation-based competency training. *Chart, 109*(3), 11-14.

Centre for Reviews and Dissemination, University of York. (2009). *Systematic reviews. CRDs guidance for undertaking reviews in healthcare.* York, UK: CRD, University of York. Retrieved from: http://www.york.ac.uk/inst/crd/pdf/Systematic_Reviews.pdf

Childs, J. C., & Sepples, S. (2006). Clinical teaching by simulation: Lessons learned from a complex patient care scenario. *Nursing Education Perspectives, 27*(3), 154-158. Retrieved from http://ezproxy.library.dal.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=c8h&AN=2009200646&site=ehost-live

Clark, M. (2006). Evaluating an Obstetrical Trauma Scenario. Clinical Simulation in Nursing Education, 2(2), e75-e77.

Clark, M. (2007a). Clinical Simulation Grading Rubric, Unpublished master's thesis, Midwestern State University, Wilson School of Nursing , Wichita Falls, Texas.

Clark, M. (2007b, February). Tool time: Lessons learned in instrument development. Poster session presented at the NLN Leadership Conference, Orlando, Florida.

College of Registered Nurses of Nova Scotia. (2012). Standards *of Practice for Registered Nurses.* Retrieved from http://www.crnns.ca/documents/RNStandards.pdf

College of Registered Nurses of Nova Scotia. (2013). Entry-level Competencies for Registered Nurses. http://crnns.ca/practice-standards/entry-level-competencies/rn-entry-level-competencies/

College of Registered Nurses of Nova Scotia. (2015). Continuing Competence Program.http://crnns.ca/practice-standards/quality-assurance/ccp/

Conn, V. S., Sang-arun, I., Rath, S., Peeranuch J., Rohini, W. & Yashodhara, D. (2003). Beyond MEDLINE for literature searches. *Journal of Nursing Scholarship, 35*(2), 177-82. Retrieved from http://search.proquest.com.ezproxy.library.dal.ca/docview/230486000?accountid=10406

Cowan, D. T., Norman, I., & Coopamah, V. P. (2007). Competence in nursing practice: A controversial concept – A focused review of literature. *Accident and Emergency Nursing, 15*(1), 20-26. doi:10.1016/j.aaen.2006.11.002

Decker, S., Sportsman, S., Puetz, L., & Billings, L. (2008). The evolution of simulation and its contribution to competency. *Journal of Continuing Education in Nursing, 39*(2), 74-80. Retrieved from http://ezproxy.library.dal.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=c8h&AN=2009798308&site=ehost-live

Doolen, J. (2012). *The development of the simulation thinking rubric.* University of Northern Colorado). , 174 p. (UMI Order AAI3555113.) Retrieved from http://ezproxy.library.dal.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=c8h&AN=2012250524&site=ehost-live. (2012250524).

EDCAN- National Cancer Nursing Education Project (2008). Australia. http://www.cancerlearning.gov.au/edcan_resources/#/xml/about

Facione, N.C., Facione, P. A., & Sanchez, C. A., MA. (1994). Critical thinking disposition as a measure of competent clinical judgment: The development of the California Critical Thinking Disposition Inventory. *Journal of Nursing Education, 33*(8), 345-350. Retrieved from http://search.proquest.com.ezproxy.library.dal.ca/docview/1026710544?accountid=10406

Facione, P.A., & Facione, N.P. (1998*). The California Critical Thinking Skills Test: CCTST test manual.* Millbrae, CA; California Academic Press.

Fahy, A., Tuohy, D., McNamara, M., C., Butler, M., P., Cassidy, I., & Bradshaw, C. (2011). Evaluating clinical competence assessment. *Nursing Standard, 25*(50), 42-48. Retrieved from http://ezproxy.library.dal.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=c8h&AN=2011253711&site=ehost-live

Frasure, J. (2008). Analysis of instruments measuring nurses' attitudes towards research utilization: A systematic review. *Journal of Advanced Nursing*, *61*(1), 5-18. doi:10.1111/j.1365-2648.2007.04525.x

Frontiero, L. & Glynn, P. (2012). Evaluation of senior nursing students' performance with high fidelity simulation. *Online Journal of Nursing Informatics*, *16*(3), Retrieved from http://ojni.org/issues/?p=2037

Gaba, D. (2004). The future vision of simulation in health care. *Quality Safety Health Care, 13*, i2-i10. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=cinref&AN=QSHC.AC.IB.GABA.FVISH

Galloway S. (2009). Simulation techniques to bridge the gap between novice and competent healthcare professionals. *Online Journal of Issues in Nursing, 14*(2), 1-9.

Gantt, L. T., & Webb-Corbett, R. (2010). Using simulation to teach patient safety behaviors in undergraduate nursing education. *Journal of Nursing Education, 49*(1), 48-51. doi:10.3928/01484834-20090918-10 &site=ehost-live

Garrett, B., MacPhee, M., & Jackson, C. (2010). High-fidelity patient simulation: Considerations for effective learning. *Nursing Education Perspectives, 31*(5), 309-313. Retrieved from http://ezproxy.library.dal.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=c8h&AN=2010862051&site=ehost-live

Garside, J. R., & Nhemachena, J. Z. Z. (2013). A concept analysis of competence and its transition in nursing. *Nurse Education Today, 33*(5), 541-545. doi:http://dx.doi.org.ezproxy.library.dal.ca/10.1016/j.nedt.2011.12.007

Goodstone, L., & Goodstone, M., S. (2013). Use of simulation to develop a medication administration safety assessment tool. *Clinical Simulation in Nursing, 9*(12), e609-15. doi:10.1016/j.ecns.2013.04.017

Grimshaw, J. (2010). *A Guide to Knowledge Synthesis. A Knowledge Synthesis Chapter*. Canadian Institutes of Health Research. Retrieved from: http://www.cihr-irsc.gc.ca/e/41382.html

Grant, J. S., Moss, J., Epps, C., & Watts, P. (2010). Using video-facilitated feedback to improve student performance following high-fidelity simulation. *Clinical Simulation in Nursing, 6*(5), e177-e184. doi:http://dx.doi.org.ezproxy.library.dal.ca/10.1016/j.ecns.2009.09.001

Gunberg Ross, J. (2012). Simulation and psychomotor skill acquisition: A review of the literature. *Clinical Simulation in Nursing*, *8*, e429-e435

Haggard, L., K. (2013). *High fidelity patient simulation and safety competencies in nursing students.* Capella University). , 130 p. (UMI Order AAI3567850.) Retrieved from http://ezproxy.library.dal.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=c8h&AN=2012354635&site=ehost-live. (2012354635).

Higgins, J.P.T., & Green, S. (Eds.). (2011) *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.1.0 The Cochrane Collaboration. Retrieved from www.cochrane-handbook.org

Holopainen, A., Hakulinen-Viitanen, T., & Tossavainen, K. (2008). Systematic review- a method for nursing research. *Nurse Researcher, 16(*1), 72-83.

Horan, K. M. (2009). Using the human patient simulator to foster critical thinking in critical situations. *Nursing Education Perspectives, 30*(1), 28-30. Retrieved from http://ezproxy.library.dal.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=c8h&AN=2010196673&site=ehost-live

Houde, S. (2009). The systematic review of the literature: A tool for evidence-based policy. *Journal of Gerontological Nursing*, *35(*9), 9-12. Doi: 10.3928/00989134-20090731-05

Houser, J. (2012). *Nursing research. Reading, using and creating evidence.* Sudbury, MA: Jones & Bartlett

Huitt, W., & Cain, S. (2005). An overview of the conative domain. *Educational Psychology Interactive*. Valdosta, GA: Valdosta State University. Retrieved [date] from http:/www.edpsycinteractive.org /brilstar/chapters/conative.pdf

Isaacson, J. J., & Stacy, A. S. (2009). Rubrics for clinical evaluation: Objectifying the subjective experience. *Nurse Education in Practice, 9*(2), 134-140. doi:10.1016/j.nepr.2008.10.015

Jeffries, P. R. (2005). A framework for designing, implementing, and evaluating: Simulations used as teaching strategies in nursing. *Nursing Education Perspectives, 26*(2), 96-103. Retrieved from http://ezproxy.library.dal.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=c8h&AN=2005104234&site=ehost-live

Jeffries, P. R. & National League for Nursing (U.S.A.) (2007*) Simulation in Nursing Education: from conceptualization to evaluation.* National League for Nursing, New York.

Jensen, R. (2013). Clinical reasoning during simulation: Comparison of student and faculty ratings. *Nurse Education in Practice, 13*(1), 23-28. doi:http://dx.doi.org.ezproxy.library.dal.ca/10.1016/j.nepr.2012.07.001

Joanna Briggs Institute. (2013). The JBI Model. Retrieved October 2013 from http://joannabriggs.org/jbi-approach.html#tabbed-nav=JBI-approach

Kardong-Edgren, S., Adamson, K.A., Fitzgerald, C. (2010) A review of currently published evaluation instruments for human patient simulation*. Clinical Simulation in Nursing*, *6*, e25-e35.

Kim, M., & Shin, M. (2013). Development and evaluation of simulation-based training for obstetrical nursing using human patient simulators. *CIN: Computers, Informatics, Nursing, 31*(2), 76-84. doi:10.1097/NXN.0b013e3182701041

Kneebone, R. L., Scott, W., Darzi, A., & Horrocks, M. (2004). Simulation and clinical practice: Strengthening the relationship. *Medical Education, 38*(10), 1095-1102. Retrieved from http://ezproxy.library.dal.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=c8h&AN=2009399773&site=ehost-live

Kuiper, R.A., Heinrich, C., Matthias, A., Graham, M., Bell-Kotwall, L. (2008). Debriefing with the OPT model of clinical reasoning during high fidelity patient simulation. *International Journal of Nursing Education Scholarship, 5*(1), Article17-Article17. doi:10.2202/1548-923X.1466

Lapkin, S., Levett-Jones, T., Bellchambers, H., & Fernandez, R. (2010). Effectiveness of patient simulation manikins in teaching clinical reasoning skills to undergraduate nursing students: A systematic review. *Clinical Simulation in Nursing, 6*(6), e207. doi:10.1016/j.ecns.2010.05.005

Lasater, K. (2005). Human patient simulation: Impact on the development of clinical judgment. *Communicating Nursing Research, 38*, 199-199. Retrieved from http://ezproxy.library.dal.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=c8h&AN=2009901883&site=ehost-live

Lasater, K. (2007). Clinical judgment development: Using simulation to create an assessment rubric. *Journal of Nursing Education, 46*(11), 496-503. Retrieved from http://ezproxy.library.dal.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=c8h&AN=2009709670&site=ehost-live

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

Lasater, K. (2011). Clinical judgement: the last frontier for evaluation. *Nurse Education and Practice, 11*(2), 86-92.

Lauder, Holland, W., Roxburgh, K., Topping, M., Watson, K., Johnson, R., … Agnieszka. (2008). Measuring competence, self-reported competence and self-efficacy in pre-registration students. *Nursing Standard, 22*(20), 35-43. doi:10.7748/ns2008.01.22.20.35.c6316

Levett-Jones, T., Hoffman, K., Dempsey, J., Jeong, S. Y., Noble, D., Norton, C. A. …Hickey, N. (2010). The 'five rights' of clinical reasoning: An educational model to enhance nursing students' ability to identify and manage clinically 'at risk' patients. *Nurse Education Today*, *30*(6), 515-520. doi:http://dx.doi.org.ezproxy.library.dal.ca/10.1016/j.nedt.2009.10.020

Liaw S.Y., Chen F.G., Klainin P., Brammer J., O'Brien A., & Samarasekera D.D. (2010). Developing clinical competency in crisis event management: An integrated simulation problem-based learning activity. *Advances in Health Sciences Education, 15*(3), 403-413. Retrieved from http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L50704851; http://dx.doi.org/10.1007/s10459-009-9208-9; http://sfxhosted.exlibrisgroup.com/dal?sid=EMBASE&issn=13824996&id=doi:10.1007%2Fs10459-009-9208-9&atitle=Developing+clinical+competency+in+crisis+event+management%3A+An+integrated+simulation+problem-based+learning+activity&stitle=Adv.+Health+Sci.+Educ.&title=Advances+in+Health+Sciences+Education&volume=15&issue=3&spage=403&epage=413&aulast=Liaw&aufirst=S.Y.&auinit=S.Y.&aufull=Liaw+S.Y.&coden=&isbn=&pages=403-413&date=2010&auinit1=S&auinitm=Y.

Liu, W., Cheon, J., & Thomas, S. (2014). Interventions on mealtime difficulties in older adults with dementia: A systematic review. *International Journal of Nursing Studies, 51*(1), 14-27. doi:10.1016/j.ijnurstu.2012.12.021

Marshall, G., & Sykes, A. E. (2011). Systematic reviews: A guide for radiographers and other health care professionals. *Radiography, 17*(2), 158-164. doi:10.1016/j.radi.2010.08.007

Masui, C. and De Corte, E. 2005. Learning to reflect and to attribute constructively as basic components of self-regulated learning. *British Journal of Educational Psychology* 75 (3): 351-372.

Meakim, C., Boese, T., Decker, S., Franklin, A.E., Gloe, D., Lioce, L., … Borum, J.C. (2013). Standards of best practice: Simulation standard 1: Terminology. *Clinical Simulation in Nursing,* 9(6S), S3-S11. doi:10.1016/j.ecns.2013.04.001

Merriman, C., D., Stayt, L., C., & 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 Simulation in Nursing, 10*(3), e119-27. doi:10.1016/j.ecns.2013.09.004

Meyer, R., Allen. (2012). *Assessment of the impact of integrated simulation on critical thinking and clinical judgment in nursing instruction.* University of North Dakota). , 102 p. (UMI Order AAI3554006.) Retrieved from http://ezproxy.library.dal.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=c8h&AN=2012250540&site=ehost-live. (2012250540).

Mikasa, A. W., Cicero, T. F., & Adamson, K. A. (2013). Outcome-based evaluation tool to evaluate student performance in high-fidelity simulation. *Clinical Simulation in Nursing, 9*(9), e361-e367. doi:http://dx.doi.org.ezproxy.library.dal.ca/10.1016/j.ecns.2012.06.001

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. (2009). Reprint--preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Physical Therapy, 89*(9), 873-880

Mokkink, L.B., Terwee, C.B., Knol, D.L., Stratford, P.W., Alonso, J., Patrick, D.L.,…de Vet, H.C.W. (2010). The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. *BMC Medical Research Methodology, 10(*22*)*

Morgan, P. J., & Cleave Hogg, D. (2002). Comparison between medical students' experience, confidence and competence. *Medical Education, 36*(6), 534-539. doi:10.1046/j.1365-2923.2002.01228.x

Moule, P., Wilford, A., Sales, R., & Lockyer, L. (2008). Student experiences and mentor views of the use of simulation for learning. Nurse Education Today, *28*(7), 790-797. doi:10.1016/j.nedt.2008.03.007

Nehring, W. M., & Lashley, F. R. (2004). Current use and opinions regarding human patient simulators in nursing education: An international survey. *Nursing Education Perspectives, 25*(5), 244-248. Retrieved from http://ezproxy.library.dal.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=c8h&AN=2004187505&site=ehost-live

Nicholson, A., Christine. (2010). *Comparison of selected outcomes based on teaching strategies that promote active learning in nursing education.* University of Iowa). , 211 p. (UMI Order AAI3409502.) Retrieved from http://ezproxy.library.dal.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=c8h&AN=2011033297&site=ehost-live. (2011033297).

Norman, J. (2012). Systematic review of the literature on simulation in nursing education. *ABNF Journal, 23*(2), 24-28. Retrieved from http://ezproxy.library.dal.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=c8h&AN=2011552819&site=ehost-live

Observable. (n.d.). In Miriam-Webster`s online dictionary. Retrieved from http://www.merriam-webster.com/

Oldenburg, N. L., Maney, C., & Plonczynski, D. J. (2013). Traditional clinical versus simulation in 1st semester clinical students: Students perceptions after a 2nd semester clinical rotation. *Clinical Simulation in Nursing, 9*(7), e235-e241. doi:http://dx.doi.org.ezproxy.library.dal.ca/10.1016/j.ecns.2012.03.006.

Patton S.K. (2013). Pilot study to evaluate consistency among raters of a clinical simulation. *Nursing Education Perspectives, 34*(3), 194-195. Retrieved from http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L369272642; http://dx.doi.org/10.5480/1536-5026-34.3.194; http://sfxhosted.exlibrisgroup.com/dal?sid=EMBASE&issn=15365026&id=doi:10.5480%2F1536-5026-34.3.194&atitle=Pilot+study+to+evaluate+consistency+among+raters+of+a+clinical+simulation&stitle=Nurs.+Educ.+Persp.&title=Nursing+Education+Perspectives&volume=34&issue=3&spage=194&epage=195&aulast=Patton&aufirst=Susan+K.&auinit=S.K.&aufull=Patton+S.K.&coden=&isbn=&pages=194-195&date=2013&auinit1=S&auinitm=K.

Pearson, A. (Ed.) *Joanna Briggs Institute Reviewers' Manual* (2014 ed.). Adelaide, Australia: Joanna Briggs Institute. Retrieved from: http://joannabriggs.org/assets/docs/sumari/ReviewersManual-2014.pdf

Pearson, A., Wiechula, R., Court, A., & Lockwood, C. (2005). The JBI model of evidence-based healthcare. *International Journal of Evidence-Based Healthcare, 3*(8), 207-215. Retrieved from http://ezproxy.library.dal.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=c8h&AN=2009128735&site=ehost-live

Petticrew, M. & Roberts, H. (2006). *Systematic reviews in the social sciences. A practical guide.* Malden, MA: Blackwell Publishing.

Polit, D., & Beck, C. (2008) *Nursing research. Generating and assessing evidence for nursing practice (*8th ed.). Philadelphia, PA: Lippincott, Williams & Wilkins.

Polkki, T., Kanste, O., Kaarianen, M., Elo, S., & Kyngas, H. (2014). The methodological quality of systematic reviews published in high-impact nursing journals: A review of the literature. *Journal of Clinical Nursing, 23*(3-4), 315-332. doi:10.1111/jocn.12132

Radhakrishnan, K., Roche, J. P., & Cunningham, H. (2007). Measuring clinical practice parameters with human patient simulation: A pilot study. *International Journal of Nursing Education Scholarship, 4*(1), 1-11. Retrieved from http://ezproxy.library.dal.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=c8h&AN=2009526959&site=ehost-live

Ravert, P. (2008). Patient simulator sessions and critical thinking. *Journal of Nursing Education*, *47*(12), 557-562. Retrieved from http://ezproxy.library.dal.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=c8h&AN=2010128016&site=ehost-live

Reeves, T. C. (2006). How do you know they are learning? : The importance of alignment in higher education. *International Journal of* ***Learning*** *Technology, 2*(4), 294–309. DOI: 10.1504/IJLT.2006.011336 retrieved from http://net.educause.edu/ir/library/pdf/eli08105a.pdf

Rhodes, M., & Curran, C. (2005). Use of the human patient simulator to teach clinical judgment skills in a baccalaureate nursing program. *CIN: Computers, Informatics, Nursing, 23*(5), 256-264. Retrieved from http://ezproxy.library.dal.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=c8h&AN=2009035867&site=ehost-live

Rosen, K. R. (2008). The history of medical simulation. *Journal of Critical Care, 23*(2), 157-166. Retrieved from http://ezproxy.library.dal.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=c8h&AN=2009955233&site=ehost-live

Schlairet, M. C., & Pollock, J. W. (2010). Equivalence testing of traditional and simulated clinical experiences: Undergraduate nursing students' knowledge acquisition. *Journal of Nursing Education, 49*, 1, 43-47.

Schuwirth, L. W., & van der Vleuten, C.P.M. (2003). The use of clinical simulations in assessment. *Medical Education, 37*, 65-71. Retrieved from http://ezproxy.library.dal.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=c8h&AN=2009389738&site=ehost-live

Seropian, M. A. (2003). General concepts in full-scale simulation: getting started. *Anesthesia and Analgesia, 9* (7), 1695-1705.

Shearer, J. E. (2013). High-fidelity simulation and safety: An integrative review. *Journal of Nursing Education, 52*(1), 39-45. doi:http://dx.doi.org.ezproxy.library.dal.ca/10.3928/01484834-20121121-01

Shelestak, D., Voshall, B. (2014). Examining validity, fidelity, and reliability of human patient simulation. *Journal of Clinical Simulation in* Nursing. *10* (5), e257-60.

Simmons, B. (2010). Clinical reasoning: Concept analysis. Journal *of Advanced Nursing, 66*(5), 1151-1158. doi:10.1111/j.1365-2648.2010.05262.

Smith, S.A. (2012). Nurse Competence: A Concept Analysis. *International Journal of Nursing Knowledge*. *23*(23)172-182

Smith, S. J., & Roehrs, C. J. (2009). High-fidelity simulation: Factors correlated with nursing student satisfaction and self-confidence. *Nursing Education Perspectives, 30*(2), 74-78. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=c8h&AN=2010258637&site=ehost-live

Sportsman, S., Bolton, C., Bradshaw, P., Close, D., Lee, M., Townley, N., & Watson, M. N. (2009). A regional simulation center partnership: Collaboration to improve staff and student competency. *Journal of Continuing Education in Nursing, 40*(2), 67-73. Retrieved from http://ezproxy.library.dal.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=c8h&AN=2010191485&site=ehost-live

Starkweather, A. R., & Kardong-Edgren, S. (2008). Diffusion of innovation: Embedding simulation into nursing curricula. *International Journal of Nursing Education Scholarship, 5*(1), 1-11. Retrieved from http://ezproxy.library.dal.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=c8h&AN=2009894635&site=ehost-live

Strickland, H. P. (2013). *Comparing lasater's clinical judgment rubric scores across faculty, self-assessment, & outcome scores.* (Ed.D., The University of Alabama). *ProQuest Dissertations and Theses,* Retrieved from).http://search.proquest.com.ezproxy.library.dal.ca/docview/1505373092?accountid=10406 (1505373092).

Swanson, E., A., Nicholson, A., C., Boese, T., A., Cram, E., Stineman, A. M., & Tew, K. (2011). Comparison of selected teaching strategies incorporating simulation and student outcomes. *Clinical Simulation in Nursing, 7*(3), e81-90. doi:10.1016/j.ecns.2009.12.011

Tanner, C. A. (2006). Thinking like a nurse: A research-based model of clinical judgment in nursing. *Journal of Nursing Education, 45*(6), 204-211. Retrieved from http://ezproxy.library.dal.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=c8h&AN=2009209758&site=ehost-live

Terwee, C. B., Mokkink, L. B., Knol, D. L., Ostelo, R. W. J. G., Bouter, L.M., de Vet, H. C. W. (2012). Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Quality of Life Research, 21*, 651-657.

Tricco, A.C., Tetzlaff, J., & Moher, D. (2011). The art and science of knowledge synthesis. *Journal of Clinical Epidemiology*, *64*(1), 11-20. Retrieved from: https://www-clinicalkey-com.ezproxy.library.dal.ca/#!/ContentPlayerCtrl/doPlayContent/1-s2.0-S0895435609003618

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. Retrieved from http://ezproxy.library.dal.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=c8h&AN=2010111785&site=ehost-live

Victor-Chmil, J., & Larew, C. (2013). Psychometric properties of the Lasater Clinical Judgment Rubric. *International Journal of Nursing Education Scholarship*, *10*, 45-52.

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

Webb, C., & Roe, B. (2007). *Reviewing research evidence for nursing practice. Systematic reviews*. Oxford, UK: Blackwell Publishing. Except Chapter 4 © 2006 by John Wiley and Sons.

Whittemore, R., & Knafl, K. (2005). The integrative review: Updated methodology. *Journal of Advanced Nursing, 52*(5), 546-553. doi:10.1111/j.1365-2648.2005.03621.x

Windle, P. E. (2010). The systematic review process: An overview. *Journal of PeriAnesthesia Nursing, 25*(1), 40-42. doi:10.1016/j.jopan.2009.12.001

Wolf, L., Dion, K., Lamoureaux, E., Kenny, C., Curnin, M., Hogan, M. A., & Cunningham, H. (2011). Using simulated clinical scenarios to evaluate student performance. *Nurse Educator, 36*(3), 128-134. doi:10.1097/NNE.0b013e31821612

Yanhua, C., & Watson, R. (2011). A review of clinical competence assessment in nursing. *Nurse Education Today. 31*(8):832-6. doi: 10.1016/j.nedt.2011.05.003. Epub 2011 Jun 1.

Yuan, H. B., Williams, B. A., & Fang, J. B. (2012). The contribution of high-fidelity simulation to nursing students' confidence and competence: A systematic review. *International Nursing Review, 59*(1), 26-33. doi:http://dx.doi.org/10.1111/j.1466-7657.2011.00964.x

Ziv, A., Wolpe, P.R., Small, S., & Glick, S. (2003). Simulation-based based medical education: An ethical imperative. *Academic Medicine, 78*(8), 783-788.

Appendix A

Definitions

Term	Definition
Clinical Judgment	The art of making a series of decisions to determine whether to take action based on various types of knowledge. The individual recognizes changes and salient aspects in a clinical situation, interprets their meaning, responds appropriately, and reflects on the effectiveness of the intervention (Meakim et al., 2013).
Clinical Reasoning	A logical process by which nurses (and other clinicians) collect cues, process the information, come to an understanding of a patient problem or situation, plan and implement interventions, evaluate outcomes, and reflect on and learn from the process" (Lapkin et al., 2010) (p e209).
Competence	A standardized requirement for an individual to properly perform a specific role. It encompasses a combination of discrete and measurable knowledge, skills and attitudes that are essential for patient safety and quality patient care (Meakim et al., 2013).
Critical Thinking	A disciplined process that requires validation of the data, including any assumptions that may influence thoughts and actions, and then careful reflection on the entire process while evaluating the effectiveness of what has been determined as the necessary action(s) to take. This process entails purposeful, goal- directed thinking and is based on scientific principles and methods (evidence) rather than assumptions or conjecture (Meakim et al., 2013).
Fidelity	Fidelity is the believability, or the degree to which a simulated experience approaches reality, .involving a variety of dimensions including: physical factors such as environment, equipment, and related tools; psychological factors such as emotions, beliefs, self-awareness; social factors such as motivation and goals; culture of group; degree of openness and trust, and modes of thinking (Meakim et al., 2013 p. S5). The three levels of sophistication are low, moderate and high. (Jeffries, 2007 p. 28)
High Fidelity Simulation	Experiences using full scale computerized patient simulators, virtual reality or standardized patients that are extremely realistic and provide a high level of interactivity and realism for the learner (Meakim et al., 2013; Jeffries, 2007).
High Fidelity Patient Simulator (HFPS)	This manikin can be programmed to breathe, speak, has palpable pulses, audible breath sounds, can react appropriately to medications, and defibrillation, and can be programmed to deliver countless scenarios (Horan, 2006). Full body simulator that can be programmed to respond to affective and psychomotor changes to simulate a variety of patient conditions. The simulator provides a high degree of realism and interactivity. The simulator allows for patient responses to questions by students, providing feedback during clinical simulation activities and

	allowing communication between the student and patient. This patient verbal communication can be accomplished through prerecorded vocal responses or with responses given by the instructor through a microphone connected to the manikin. (Jeffries, 2007) High-fidelity human patient simulators also have the ability to simulate electrocardiogram (EKG/ECG) readings and other patient information on monitors, such as those found on critical care units (Jeffries, 2007).
Human patient simulator	Same as high fidelity patient simulator.
Observational	Adjective Based on observation or experience <her reports on the great apes were based on firsthand *observational* evidence> **http://www.merriamwebster.com/thesaurus/observational?show=0&t=1391378816**
Observable	Adjective capable of being seen <scientists often work with phenomena that are not directly *observable*>http://www.merriam-webster.com/thesaurus/observable
Outcomes	The measurable results of the participant's progress toward meeting a set of objectives. Expected outcomes are the change in knowledge, skills and attitude (KSA) as a result of the simulated experience (Meakim et al., 2013).
Reliability of measuring instruments	The degree of consistency or stability with which an instrument measures an attribute (Polit & Beck, 2008 p. 764). Reliability has three key attributes: stability, internal consistency and equivalence. Stability of an instrument is the extent to which similar results are obtained on repeated administrations. This can be determined by test - re-test method. Internal consistency or homogeneity is the extent to which all the instrument's items measure the same attribute. Cronbach's alpha is one method most commonly used as an index of internal consistency to estimate the extent to which different subparts of an instrument are reliably measuring the critical attribute (Polit & Beck, 2008 p. 455). Equivalence concerns the degree to which two or more independent observers agree about the scoring on an instrument. This is referred to as inter-rater reliability. When there is a high level of agreement it is assumed that measurement errors have been minimized (Polit & Beck, 2008 p. 455).
Validity	The degree to which a test or evaluation tool accurately measures the intended concept of interest. Aspects of validity are: face validity, content validity, criterion-related validity, and construct validity (Polit & Beck, 2008). This study will focus on all validity measures of the instruments used to assess student competency outcomes following high fidelity simulation-based clinical experiences.

	Face validity refers to whether the instrument appears, on the face of it, as though it is measuring the appropriate construct. Content validity concerns the degree to which the instrument has an appropriate sample of items for the construct being measured to adequately cover the construct domain (Polit & Beck, 2008 p. 458). It is relevant for affective and cognitive measures. It may be determined by a panel of substantive experts who evaluate and document the content validity for the tool or by calculating the content validity index (Polit & Beck, 2008 p. 459). Criterion- related validity involves determining the relationship between an instrument and an external criterion. The three types of criterion-related validity are predictive validity, concurrent validity and discriminate validity. Predictive validity refers to the adequacy of the instrument in differentiating between people's performance on a future criterion. Concurrent validity refers to the instrument's ability to distinguish individuals who differ on a present criterion (Polit & Beck, 2008 p. 460). Discriminate validity refers to the tool's ability to discriminate between those who have a characteristic from those who don't (Houser, 2012). Construct validity concerns the degree to which the instrument measures the construct under question (Polit & Beck, 2008 p. 750). It is the validity of inferences from observed persons, settings and interventions in a study to the constructs that these instances may represent. It is primarily validated by testing a hypothesis linked to a theoretical perspective regarding the construct (Polit & Beck, 2008 p. 461).

Appendix B

PRISMA Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(6): e1000097. doi:10.1371/journal.pmed1000097

Appendix C

Study Selection Form

Study Selection Form for SR of SBCE and measurement instruments of competence outcomes				
Bibliographic details of study:				
PICOT & limits	**Inclusion Criteria**	**Yes**	**No**	**Undecided**
Time frame	Study done in years 2000-2014			
Language	English, already translated into English			
Setting	High fidelity simulation lab			
Participants	Undergraduate nursing students Associate degree Baccalaureate degree in any year of their program			
Intervention	High fidelity simulation-based clinical experience includes: - high- fidelity computerized patient simulator -clinical scenario with realistic environment and equipment			
Type of study	**Quantitative:** Experimental: -randomized controlled trial Quasi-experimental: -non-randomized controlled trial -comparative design before-and-after study - interrupted time series Pilot study of experimental or quasi-experimental studies Observational: - pre and post cohort studies -instrument development study			
Outcomes	observational measurement instrument -name student's competence outcomes: -CT -CJ -CR -competence			
Comparative Intervention	Clinical practicum on-site at a healthcare agency or community site. -Lab-based clinical experiences -Clinical scenario case study -Clinical scenario with standardized patients -Lecture classes on clinical scenarios -Low or medium fidelity simulations -Low or medium clinical skill simulation experiences.			
Action	with rational (yes, no, or undecided for phase 1)			
Version 2 03/05/14 BB				

Appendix D Study Appraisal Checklist

The COSMIN checklist

Contact
CB Terwee, PhD
VU University Medical Center
Department of Epidemiology and Biostatistics
EMGO Institute for Health and Care Research
1081 BT Amsterdam
The Netherlands
Website: www.cosmin.nl, www.emgo.nl
E-mail: cb.terwee@vumc.nl

Step 1. Evaluated measurement properties in the article

	Internal consistency	Box A
	Reliability	Box B
	Measurement error	Box C
	Content validity	Box D
	Structural validity	Box E
	Hypotheses testing	Box F
	Cross-cultural validity	Box G
	Criterion validity	Box H
	Responsiveness	Box I
	Interpretability	Box J

Step 2. Determining if the statistical method used in the article are based on CTT or IRT

Box General requirements for studies that applied Item Response Theory (IRT) models

		yes	no	?
1	Was the IRT model used adequately described? e.g. One Parameter Logistic Model (OPLM), Partial Credit Model (PCM), Graded Response Model (GRM)	□	□	
2	Was the computer software package used adequately described? e.g. RUMM2020, WINSTEPS, OPLM, MULTILOG, PARSCALE, BILOG, NLMIXED	□	□	
3	Was the method of estimation used adequately described? e.g. conditional maximum likelihood (CML), marginal maximum likelihood (MML)	□	□	
4	Were the assumptions for estimating parameters of the IRT model checked? e.g. unidimensionality, local independence, and item fit (e.g. differential item functioning (DIF))	□	□	□

Step 3. Determining if a study meets the standards for good methodological quality

Box A. Internal consistency

		yes	**no**	**?**
1	Does the scale consist of effect indicators, i.e. is it based on a reflective model?	□	□	□
Design requirements		**yes**	**no**	**?**
2	Was the percentage of missing items given?	□	□	
3	Was there a description of how missing items were handled?	□	□	
4	Was the sample size included in the internal consistency analysis adequate?	□	□	□
5	Was the unidimensionality of the scale checked? i.e. was factor analysis or IRT model applied?	□	□	
6	Was the sample size included in the unidimensionality analysis adequate?	□	□	□
7	Was an internal consistency statistic calculated for each (unidimensional) (sub)scale separately?	□	□	□
8	Were there any important flaws in the design or methods of the study?	□	□	
Statistical methods		**yes**	**no**	**NA**
9	for Classical Test Theory (CTT): Was Cronbach's alpha calculated?	□	□	□
10	for dichotomous scores: Was Cronbach's alpha or KR-20 calculated?	□	□	□
11	for IRT: Was a goodness of fit statistic at a global level calculated? e.g. χ^2, reliability coefficient of estimated latent trait value (index of (subject or item) separation)	□	□	□

Box B. Reliability: relative measures (including test-retest reliability, inter-rater reliability and intra-rater reliability)

Design requirements		**yes**	**no**	**?**
1	Was the percentage of missing items given?	□	□	
2	Was there a description of how missing items were handled?	□	□	
3	Was the sample size included in the analysis adequate?	□	□	□
4	Were at least two measurements available?	□	□	
5	Were the administrations independent?	□	□	□
6	Was the time interval stated?	□	□	
7	Were patients stable in the interim period on the construct to be measured?	□	□	□

8	Was the time interval appropriate?	□	□	□
9	Were the test conditions similar for both measurements? e.g. type of administration, environment, instructions	□	□	□
10	Were there any important flaws in the design or methods of the study?	□	□	

	Statistical methods	**yes**	**no**	**NA**	**?**
11	for continuous scores: Was an intraclass correlation coefficient (ICC) calculated?	□	□	□	
12	for dichotomous/nominal/ordinal scores: Was kappa calculated?	□	□	□	
13	for ordinal scores: Was a weighted kappa calculated?	□	□	□	□
14	for ordinal scores: Was the weighting scheme described? e.g. linear, quadratic	□	□	□	

Box C. Measurement error: absolute measures

	Design requirements	**yes**	**no**	**?**
1	Was the percentage of missing items given?	□	□	
2	Was there a description of how missing items were handled?	□	□	
3	Was the sample size included in the analysis adequate?	□	□	□
4	Were at least two measurements available?	□	□	
5	Were the administrations independent?	□	□	□
6	Was the time interval stated?	□	□	
7	Were patients stable in the interim period on the construct to be measured?	□	□	□
8	Was the time interval appropriate?	□	□	□
9	Were the test conditions similar for both measurements? e.g. type of administration, environment, instructions	□	□	□
10	Were there any important flaws in the design or methods of the study?	□	□	

	Statistical methods	**yes**	**no**	**?**
11	for CTT: Was the Standard Error of Measurement (SEM), Smallest Detectable Change (SDC) or Limits of Agreement (LoA) calculated?	□	□	

Box D. Content validity (including face validity)

	General requirements	yes	no	?
1	Was there an assessment of whether all items refer to relevant aspects of the construct to be measured?	□	□	□
2	Was there an assessment of whether all items are relevant for the study population? (e.g. age, gender, disease characteristics, country, setting)	□	□	□
3	Was there an assessment of whether all items are relevant for the purpose of the measurement instrument? (discriminative, evaluative, and/or predictive)	□	□	□
4	Was there an assessment of whether all items together comprehensively reflect the construct to be measured?	□	□	□
5	Were there any important flaws in the design or methods of the study?	□	□	

Box E. Structural validity

		yes	no	?
1	Does the scale consist of effect indicators, i.e. is it based on a reflective model?	□	□	□
	Design requirements	**yes**	**no**	**?**
2	Was the percentage of missing items given?	□	□	
3	Was there a description of how missing items were handled?	□	□	
4	Was the sample size included in the analysis adequate?	□	□	□
5	Were there any important flaws in the design or methods of the study?	□	□	
	Statistical methods	**yes**	**no**	**NA**
6	for CTT: Was exploratory or confirmatory factor analysis performed?	□	□	□
7	for IRT: Were IRT tests for determining the (uni-) dimensionality of the items performed?	□	□	□

Box F. Hypotheses testing

	Design requirements	yes	no	?
1	Was the percentage of missing items given?	□	□	
2	Was there a description of how missing items were handled?	□	□	
3	Was the sample size included in the analysis adequate?	□	□	□

4	Were hypotheses regarding correlations or mean differences formulated a priori (i.e. before data collection)?	□	□	□
		yes	**no**	**NA**
5	Was the expected *direction* of correlations or mean differences included in the hypotheses?	□	□	□
6	Was the expected absolute or relative *magnitude* of correlations or mean differences included in the hypotheses?	□	□	□
7	for convergent validity: Was an adequate description provided of the comparator instrument(s)?	□	□	
8	for convergent validity: Were the measurement properties of the comparator instrument(s) adequately described?	□	□	
9	Were there any important flaws in the design or methods of the study?	□	□	
Statistical methods		**yes**	**no**	**NA**
10	Were design and statistical methods adequate for the hypotheses to be tested?	□	□	□

Box G. Cross-cultural validity				
Design requirements		**yes**	**no**	**?**
1	Was the percentage of missing items given?	□	□	
2	Was there a description of how missing items were handled?	□	□	
3	Was the sample size included in the analysis adequate?	□	□	□
4	Were both the original language in which the HR-PRO instrument was developed, and the language in which the HR-PRO instrument was translated described?	□	□	
5	Was the expertise of the people involved in the translation process adequately described? e.g. expertise in the disease(s) involved, expertise in the construct to be measured, expertise in both languages	□	□	
6	Did the translators work independently from each other?	□	□	□
7	Were items translated forward and backward?	□	□	□
8	Was there an adequate description of how differences between the original and translated versions were resolved?	□	□	
9	Was the translation reviewed by a committee (e.g. original developers)?	□	□	
10	Was the HR-PRO instrument pre-tested (e.g. cognitive interviews) to check interpretation, cultural relevance of the translation, and ease of comprehension?	□	□	

11	Was the sample used in the pre-test adequately described?	□	□	
12	Were the samples similar for all characteristics except language and/or cultural background?	□	□	□
13	Were there any important flaws in the design or methods of the study?	□	□	
Statistical methods		**yes**	**no**	**NA**
14	for CTT: Was confirmatory factor analysis performed?	□	□	□
15	for IRT: Was differential item function (DIF) between language groups assessed?	□	□	□

Box H. Criterion validity

Design requirements		**yes**	**no**	**?**
1	Was the percentage of missing items given?	□	□	
2	Was there a description of how missing items were handled?	□	□	
3	Was the sample size included in the analysis adequate?	□	□	□
4	Can the criterion used or employed be considered as a reasonable ‘gold standard’?	□	□	□
5	Were there any important flaws in the design or methods of the study?	□	□	
Statistical methods		**yes**	**no**	**NA**
6	for continuous scores: Were correlations, or the area under the receiver operating curve calculated?	□	□	□
7	for dichotomous scores: Were sensitivity and specificity determined?	□	□	□

Box I. Responsiveness

Design requirements		**yes**	**no**	**?**
1	Was the percentage of missing items given?	□	□	
2	Was there a description of how missing items were handled?	□	□	
3	Was the sample size included in the analysis adequate?	□	□	□
4	Was a longitudinal design with at least two measurement used?	□	□	
5	Was the time interval stated?	□	□	
6	If anything occurred in the interim period (e.g. intervention, other relevant events), was it adequately described?	□	□	

7	Was a proportion of the patients changed (i.e. improvement or deterioration)?	□	□	
Design requirements for hypotheses testing		**yes**	**no**	**?**
For constructs for which a gold standard was not available:				
8	Were hypotheses about changes in scores formulated a priori (i.e. before data collection)?	□	□	□
		yes	**no**	**NA**
9	Was the expected *direction* of correlations or mean differences of the change scores of HR-PRO instruments included in these hypotheses?	□	□	□
10	Were the expected absolute or relative *magnitude* of correlations or mean differences of the change scores of HR-PRO instruments included in these hypotheses?	□	□	□
11	Was an adequate description provided of the comparator instrument(s)?	□	□	
12	Were the measurement properties of the comparator instrument(s) adequately described?	□	□	
13	Were there any important flaws in the design or methods of the study?	□	□	
Statistical methods		**yes**	**no**	**NA**
14	Were design and statistical methods adequate for the hypotheses to be tested?	□	□	□
Design requirement for comparison to a gold standard		**yes**	**no**	**?**
For constructs for which a gold standard was available:				
15	Can the criterion for change be considered as a reasonable gold standard?	□	□	□
16	Were there any important flaws in the design or methods of the study?	□	□	
Statistical methods		**yes**	**no**	**NA**
17	for continuous scores: Were correlations between change scores, or the area under the Receiver Operator Curve (ROC) curve calculated?	□	□	□
18	for dichotomous scales: Were sensitivity and specificity (changed versus not changed) determined?	□	□	□

Box J. Interpretability

		yes	**no**	**?**
1	Was the percentage of missing items given?	□	□	
2	Was there a description of how missing items were handled?	□	□	

3	Was the sample size included in the analysis adequate?	□	□	□
4	Was the distribution of the (total) scores in the study sample described?	□	□	
5	Was the percentage of the respondents who had the lowest possible (total) score described?	□	□	
6	Was the percentage of the respondents who had the highest possible (total) score described?	□	□	
7	Were scores and change scores (i.e. means and SD) presented for relevant (sub) groups? e.g. for normative groups, subgroups of patients, or the general population	□	□	
8	Was the minimal important change (MIC) or the minimal important difference (MID) determined?	□	□	
9	Were there any important flaws in the design or methods of the study?	□	□	

Step 4: Determining the Generalisability of the results

Box Generalisability

			yes	**no**	**NA**
		Was the sample in which the HR-PRO instrument was evaluated adequately described? In terms of:			
	1	median or mean age (with standard deviation or range)?	□	□	
	2	distribution of sex?	□	□	
	3	important disease characteristics (e.g. severity, status, duration) and description of treatment?	□	□	□
	4	setting(s) in which the study was conducted? e.g. general population, primary care or hospital/rehabilitation care	□	□	
	5	countries in which the study was conducted?	□	□	
	6	language in which the HR-PRO instrument was evaluated?	□	□	
7		Was the method used to select patients adequately described? e.g. convenience, consecutive, or random	□	□	
			yes	**no**	**?**
8		Was the percentage of missing responses (response rate) acceptable?	□	□	□

Appendix D

COSMIN Scoring System

COSMIN

COSMIN checklist with 4-point scale

Please access the scale here as it will not download for preview:

http://www.cosmin.nl/images/upload/files/COSMIN%20checklist%20with%204-point%20scale%2022%20juni%202011.pdf

Instructions

This version of the COSMIN checklist is recommended for use in systematic reviews of measurement properties. With this version it is possible to calculate overall methodological quality scores per study on a measurement property. A methodological quality score per box is obtained by taking the lowest rating of any item in a box (‘worse score counts’). For example, if for a reliability study one item in the box ‘Reliability’ is scored poor, the methodological quality of that reliability study is rated as poor. The Interpretability box and the Generalizability box are mainly used as data extraction forms. We recommend to use the Interpretability box to extract all information on the interpretability issues described in this box (e.g. norm scores, floor-ceiling effects, minimal important change) of the instruments under study from the included articles. Similar, we recommend to use the Generalizability box to extract data on the characteristics of the study population and sampling procedure. Therefore no scoring system was developed for these boxes.

Appendix E

SR Data Extraction Form of SBCE and Competence Measurement

Date of data extraction:	**yes**	**no**	**Not reported**	**Pg.**
Reviewer:				
Bibliographic details of study: *Country of origin for the study*:				
Source:				
Notes:				
Purpose of study:				
Study design:				
Population:				
Sample size:				
Setting:				
Intervention:				
Comparative intervention:				
Outcome measurement instrument:				
Outcomes: *CT* *CR* *CJ* *competence*				
Domain of learning captured: *Cognitive* *Affective* *Psychomotor* *Conative*				
Statistical analyses reported: Type-				
Reported validity of instrument: *Face-* *Content-* expert panel, literature review, qualitative study, CVI *Criterion-related*: concurrent, predictive, discriminate; correlation coefficient value, gold standard criterion *Construct-* Factor analysis for subscale; re-evaluation with new setting, problem, population; repeated use of tool.				
Reported reliability of instrument: -*Stability reliability*: test-retest correlation coefficient -*Internal consistency reliability*: Cronbach's alpha coefficient value -*Equivalence reliability*: -IRR with dichotomous data - percentage agreement or Cohen's kappa value -ICC with ordinal, interval, or ratio data- kappa value				
Version 3 03/05/14 BB				

Appendix F SR Results

Table 3
Structure of Study SR

Identification stage:

Primary reviewer (PR) formulated an answerable research question:

a. PR developed a detailed SR protocol framed with the study PICOS elements
b. PR developed study objectives, eligibility criteria for studies for SR inclusion, search strategy, eligibility criteria.
c. PR developed method of: recording search results; appraising the quality and appropriateness of the primary studies, the specific data for extraction, data extraction form, syntheses of the evidence with use of tables, and method of preparation of the research report.
d. PR identified the scale for quality assessment, and method of resolving disagreements between reviewers.

2. **Search / Selection stage:**
   a. PR and SR librarian searched the literature for relevant studies using the search terms identified in the protocol and suited to the particular database.
   b. PR and second reviewer (CS) saved search results in an electronic format (Excel database) providing a comprehensive list of studies that might meet the criteria for inclusion in the review.
   c. SR reviewers (BB, CS) systematically sorted through database titles and abstracts of all the articles retrieved by the search, and excluded irrelevant studies. Excluded studies were done on consensus basis and where there was disagreement the thesis supervisor made the final decision.
   d. SR reviewers (BB, LA) read full text of 35 articles to determine their eligibility for study inclusion. Studies were excluded on consensus basis; where there was disagreement the thesis supervisor made the final decision.
   e. PR maintained a record of each abstract and article reviewed, providing documentation of the study selection process with detailed reasons for exclusion of studies that were 'near-misses' (CRD, 2009).
3. **Data collection /Appraisal stage:**
   a. SR reviewers pilot tested the data extraction form.
   b. Data extracted by both reviewers examining each of the included primary articles and extracting relevant data to the data extraction form.
   c. Both reviewers critically appraised the final list of included studies for risk of bias through the application of the COSMIN instrument of quality appraisal.
   d. Reviewers compared COSMIN study scores for each sub category and reached consensus to ensure the results were free from bias.
4. **Analysis / Synthesis stage:**
   a. PR reviewed data extracted from the studies, and organized data into tables according to the objectives of the SR.
   b. PR reported the search strategy and results, the number of articles retrieved and number of articles rejected by criteria reason.
   c. PR analyzed data from each table to report findings of each study according to the objectives of the SR
   d. PR synthesized finding to provide an analysis of the relationships within and between studies and an overall assessment of the robustness of the evidence (CRD)
   e. PR reported the various stages of the process, and discussion of the findings to enable the reader to interpret
   f. the findings and evaluate the validity and process
   of the review. (p.48) (CRD)

Appendix F SR Results

Table 4

Descriptive Summary of Study Characteristics

SR Study ID Number Study Title Authors Year	**Type of Publication**	**Country of instrument use**	**Study Design**	**Population**	**Sample Size N=**	**Intervention Method**	**Comparator Intervention**
#1 *Competency Assessment in Simulated Response to Rescue Events*. Aronson, B; Glynn, B; Squires, T. (2012).	Journal article	USA	Non-experimental instrument development study	Sr. BN	N=152 phase 1=76 phase 2=76	Dyad participation in a cardiac patient deterioration scenario scored by 2 -3 raters	None
#2 Simulation *Evaluation using a modified Lasater Clinical Judgment Rubric.* Ashcraft, A; Opton, L; Bridges, R; Caballero, S; Veesart, A; Weaver, C. (2013)	Journal article	USA	Non-experimental descriptive study	Sr. BN	N=188 phase 1=86 phase2=102	4 different clinical scenario groups with different diagnoses	None

SR Study ID Number **Study Title** **Authors** **Year**	**Type of Publication**	**Country of instrument use**	**Study Design**	**Population**	**Sample Size** **N=**	**Intervention Method**	**Comparator Intervention**
#3 *The Development of the Simulation Thinking Rubric.* Doolen, J. (2012)	PhD Dissertation	USA	Non-experimental methodological study	1st and 4th semester BN	N=44 1st semester=22 4th semester=22	High fidelity clinical simulation scenario to demonstrate language and behavioral characteristics of higher order thinking	None
#4 *Using the Clark Simulation Evaluation Rubric with Associate Degree and Baccalaureate Nursing Students.* Gantt L.T. (2010)	Journal article	USA	Quasi-experimental pilot study	1st yr. AD and Sr. BN	N=178 AD=69 BN=109	AD evaluated in high fidelity obstetrical/ child scenarios BN evaluated in high fidelity complex medical-surgical scenarios	None
#5 *Use of Simulation to Develop a Medication Administration Safety Assessment Tool-* Goodstone,L & Goodstone, M. (2013)	Journal article	USA	Quasi-experimental pilot study posttest only	PN, AD BN	N=14	Medication admin in high fidelity simulation	None
#6 *High fidelity patient simulation and safety competencies in nursing students*. Haggard, L(2013)	PhD dissertation	USA	Non-experimental correlational design	Jr. and Sr. BN from 2 different programs	N=54	Generic 20 minute HPS simulation fundamental level pertinent for any level of nursing student	None

SR Study ID Number **Study Title** **Authors** **Year**	**Type of Publication**	**Country of instrument use**	**Study Design**	**Population**	**Sample Size** **N=**	**Intervention Method**	**Comparator Intervention**
#7 *Clinical reasoning during simulation: comparison of student and faculty ratings*. Jensen, R. (2013).	Journal article	USA	Quasi experimental descriptive study	AD and BN	N=88: semester 1= 31 AD & 7 BN semester 2= 31 AD & 19 BN	Emergent patient situations during high fidelity simulation scenarios	None
#8 *Development and Evaluation of Simulation-Based Training for Obstetrical Nursing Using Human Patient Simulators* Kim, M & Shin, M. (2013).	Journal article	South Korea	Quasi-experimental pretest-posttest	BN	N=138	High and low fidelity birthing scenario	None
#9 *Clinical judgment development: using simulation to create an assessment rubric* Lasater, K. (2007).	Journal article	USA	Exploratory mixed methods- a qualitative-quantitative-qualitative design	Jr. BN	N=26	HPS simulation as primary nurse in evolving clinical situation	None

SR Study ID Number Study Title Authors Year	Type of Publication	Country of instrument use	Study Design	Population	Sample Size N=	Intervention Method	Comparator Intervention
#10 *The Impact of High Fidelity Simulation On the Development of Clinical Judgment In Nursing Students : An Exploratory Study* Lasater, K. (2005)	PhD dissertation	USA	Exploratory qualitative and quantitative methods study	Jr. BN	N=73 development phase=47 primary nurse role exam =26	High fidelity complex care clinical simulation experience for ½ day and 1 clinical practicum day	2 clinical practicum days
#11 *Developing Clinical Competency in Crisis Event Management: An Integrated Simulation Problem-Based Learning Activit*y. Liaw et al. (2010).	Journal article	Singapore	Quasi-experimental cross-over intervention posttest design	1st yr. BN	N=63 respiratory distress=30 cardiac scenario=33	Simulation of crisis event with (SPBD)	(PBD)
#12 *Comparing the Effectiveness of Clinical Simulation versus Didactic Methods to Teach Undergraduate Adult Nursing Students to Recognize and Assess the Deteriorating Patient.* Merriman, C; Stayt, L; Ricketts, B. (2014)	Journal article	UK	Experimental randomized, controlled trial	1st yr. BN	N=33 intervention group=15 control group=19	High- fidelity clinical simulation lab experience	Classroom-based teaching

SR Study ID Number Study Title Authors Year	**Type of Publication**	**Country of instrument use**	**Study Design**	**Population**	**Sample Size N=**	**Intervention Method**	**Comparator Intervention**
#13 *Assessment of the Impact of Integrated Simulation on Critical Thinking and Clinical Judgment in Nursing Instruction* Meyer, R. (2012)	PhD dissertation	USA	Quasi - experimental non-randomized controlled trial	Sr. BN	N=22 fall semester=15 spring semester=7	3 hr. didactive instruction paired with 1 hr. simulation in fall semester	3 hr. instruction with lecture, case studies and videos in spring semester
#14 *Outcome-Based Evaluation Tool to Evaluate Student Performance in High-Fidelity Simulations* Mikasa, A; Cicero, T; Adamson, K. (2013)	Journal article	USA	Non-experimental instrument development study	BN	N=84	HPS simulated clinical event	None
#15 *Comparison of selected outcomes based on teaching strategies that promote active learning in nursing education* Nicholson, A. (2010)	PhD Dissertation	USA	Experimental posttest-only design	semester 2 BN	N=74 simulation teaching strategy =25 simulation with narrative pedagogy =27 case-based learning=22	Simulation teaching strategy simulation with narrative pedagogy	Case-based learning
#16 *A Pilot Study to Evaluate Consistency Among Raters of a Clinical Simulation* Patton, S. (2013)	Journal article	USA	Non-experimental descriptive study	Sr. BN final semester	N=24	Simulated clinical experience	None

SR Study ID Number Study Title Authors Year	Type of Publication	Country of instrument use	Study Design	Population	Sample Size N=	Intervention Method	Comparator Intervention
#17 *Measuring Clinical Practice Parameters with Human Patient Simulation: A Pilot Study* Radhakrishnan, K; Roche, J; Cunningham, H. (2007)	Journal Article	USA	Quasi-experimental 2 group posttest pilot study	Sr. BN completing a second degree	N=12 intervention=6 control =6	Two patient HPS simulation assignment with complex diagnoses: one develops a medical emergency and regular clinical practice	Regular clinical practice
#18 *Comparing Lasater's Clinical Judgment Rubric Scores Across Faculty, Self-Assessment, & Outcome Scores* Strickland, H. (2013)	PhD Dissertation	USA	Experimental randomized pretest/post-test design	BN	N=94 Experimental=48 control= 46	Cardiovascular HPS simulated clinical experience	Regular course work
#19 *Comparison of Selected Teaching Strategies Incorporating Simulation and Student Outcomes.* Boese, E; Nicholson, E; Stineman, A; Tew, K. (2011)	Journal Article	USA	Experimental posttest only design	BN	N=144	HPS simulation or HPS simulation with narrative pedagogy	Case study

Abbreviations Note: Jr. =junior; Sr. = senior; BN=baccalaureate nursing students; AD= associate degree nursing student; PN= Practical Nurse; HPS=Human Patient Simulator

Appendix F SR Results

Table 5
Descriptive Summary of Instruments' Measurable Outcomes:
CT CJ CR Competence

Study ID number / Author	Instrument type	Instrument name	Domain of learning	Type of SBCE scenario	Type of outcomes for Critical thinking CT	Type of outcomes for Clinical judgment CJ	Type of outcomes for Clinical reasoning CR	Type of outcomes for Competence
#1 Aronson, Glynn, Squires (2012)	Checklist	Heart Failure Simulation Competency Evaluation Tool (HFSCET)	Psychomotor Cognitive	Adult Cardiac Care	1,2,3,4	N2,R2,3,4	1,2,3,4,6	1,2,3,4
#2 Ashcraft et al. (2013)	Rubric	Modified Lasater Clinical Judgment Rubric (LCJR)	Psychomotor Cognitive	Chronic renal failure, congestive heart failure, diabetic keto-acidosis, myocardial infarction	1,2,3	N1,N2,N3 R1,R2,4	1,3	1,2,3,4
#3 Doolen (2012)	Rubric	Simulation Thinking Rubric	Cognitive Psychomotor	Adult: abnormal vital signs, dyspnea, pain	1,2	N1,N2,N3, R1, R2, 4	1,5,6	1,2,4,5

Study ID number / Author	Instrument type	Instrument name	Domain of learning	Type of SBCE scenario	Type of outcomes for Critical thinking CT	Type of outcomes for Clinical judgment CJ	Type of outcomes for Clinical reasoning CR	Type of outcomes for Competence
#4 Gantt (2010)	Rubric	Clark Simulation Evaluation Rubric	Cognitive Psychomotor	Obstetric/ child scenarios and complex medical-surgical scenarios	1,2,3,4		1,2,5,6	1,2,4
#5 Goodstone & Goodstone (2013)	Checklist	Medication Administration Safety Assessment Tool MASAT	Cognitive Psychomotor	Medication administration scenario	3	3	5	3,4,5
#6 Haggard (2013)	Rubric	Sweeney-Clark Simulation Performance Rubric: Haggard Modification	Cognitive Psychomotor	Fundamental level pertinent for any level of nursing student	1,2,4	R1,R2,5	1,2	1,2,4
#7 Jensen (2013)	Rubric	Lasater Clinical Judgment Rubric (LCJR)	Cognitive, Psychomotor	Emergent patient situations Adult		N1,N2,N3, R1, R2,4	1,3,6	1,4

Study ID number / Author	Instrument type	Instrument name	Domain of learning	Type of SBCE scenario	Type of outcomes for Critical thinking CT	Type of outcomes for Clinical judgment CJ	Type of outcomes for Clinical reasoning CR	Type of outcomes for Competence
#8 Kim & Shin (2013)	Checklist	A 15 item checklist developed to evaluate student skills and attitudes in each of the different scenarios	Cognitive, Psychomotor	Birthing scenario	1,2,3,4	4	3	2,3,4
#9 Lasater (2007)	Rubric	Lasater Clinical Judgment Rubric	Cognitive, Psychomotor	Evolving clinical situation	1,2,3	N1,N2,N3 R1,R2,4,5	1,3,5,6	1,2,3
#10 Lasater (2005)	Rubric	Lasater Clinical Judgment in Simulation Rubric (LCJSR)	Cognitive, Psychomotor	Complex care	1,2,3	N1,N2,N3, R1,R2, 4	1,3,6	1,2,3,4

Study ID number / Author	**Instrument type**	**Instrument name**	**Domain of learning**	**Type of SBCE scenario**	**Type of outcomes for Critical thinking CT**	**Type of outcomes for Clinical judgment CJ**	**Type of outcomes for Clinical reasoning CR**	**Type of outcomes for Competence**
#11 Liaw et al. (2010)	Checklist	Two sets of Checklists developed by researcher	Cognitive, Psychomotor	Care of patients respiratory & cardiovascular disorders	1,2,3	N2,4	1,2,3,6	2,3,4
#12 Merriman et al. (2014)	Checklist	Checklist of 24 performance criteria	Cognitive, Psychomotor	Adult emergency Deteriorating patient situation	1,2,3	N1, N2, 4	1,2,3,4,6	1,3,4
#13 Meyer (2012)	Rubric	Modified Lasater Clinical Judgment Rubric	Cognitive Psychomotor	Adult Complex care	1,2,3	N1,N2,N3, R2,4	1,2,6	1,2,3,4,5
#14 Mikasa, Cicero, Adamson (2013)	Rubric	The Seattle University Simulation Evaluation Tool	Cognitive Psychomotor	Not specified	1,2,3,4	R2,4,5	2,3,4,5.6	2,3,4

Study ID number / Author	**Instrument type**	**Instrument name**	**Domain of learning**	**Type of SBCE scenario**	**Type of outcomes for Critical thinking CT**	**Type of outcomes for Clinical judgment CJ**	**Type of outcomes for Clinical reasoning CR**	**Type of outcomes for Competence**
#15 Nicholson (2010)	Rubric	Student Performance Demonstration Rubric	Cognitive Psychomotor	Adult Cardiac situation	1,2,3	N1,N2,N3, R2,4	1,2,3,4,5,6	1,2,3,4,5
#16 Patton (2013)	Checklist	The Creighton Simulation Evaluation Instrument (C-SEI)	Cognitive Psychomotor	Not specified	1,2,3	R2,4	3,6	2,3
#17 Radhakrishnan, Roche, Cunningham (2007)	Checklist	Clinical Simulation Evaluation Tool (CSET) faculty-developed	Cognitive Psychomotor	Complex diagnoses and one of which goes on to develop a medical emergency	1,2,3	N1,R2,3,4	3,4,5,6.	1,2,3,4

Study ID number / Author	**Instrument type**	**Instrument name**	**Domain of learning**	**Type of SBCE scenario**	**Type of outcomes for Critical thinking CT**	**Type of outcomes for Clinical judgment CJ**	**Type of outcomes for Clinical reasoning CR**	**Type of outcomes for Competence**
#18 Strickland (2013)	Rubric	Lasater Clinical Judgment Rubric (LCJR)	Cognitive Psychomotor	Cardiovascular HPS	1,2,3	N1,N2,N3, R1,R2,3,4,5	1,3,4,5,6	1,2,3,4,5
#19 Boese, Nicholson, Stineman, Tew (2011)	Rubric	Student Performance Demonstration Rubric	Cognitive, Psychomotor		1,3	N2,4	1,3,6	1,3,4

Measurable Outcomes: CJ CT CR Competence:

COMPETENCE PROCESS OUTCOMES: 1 recognizing pt. deviations, 2 communication, 3 technical skills, 4 skill performance 5 information seeking

CT PROCESS OUTCOMES: 1 assessment skills, 2 communication, 3 technical skills, 4 explanation

CR PROCESS OUTCOMES : 1 identifying deteriorating pt. status; 2 describe the patient situation, 3 clinical skill performance, , 4 collect new patient information, 5 review information, 6 choose a course of action

CJ PROCESS OUTCOMES: **noticing**; N1 focused observation, N2 recognizing deviations from expected patterns, N3 information seeking;
responding: R1 calm, competent manner, R2 clear communication***; 3*** safety; 4 interventions; 5 delegation

Appendix F SR Results

Table 6

Studies Categorized by Scenario Type

OBS Maternal/Child	Adult Medical Surgical	Adult Cardiology	Scenario Not Defined
#4 obstetrical/child scenarios	#2 chronic renal failure, diabetic keto-acidosis	# 1 cardiac	#9 evolving clinical situation
#8 birthing scenario; pregnancy in different stages	# 3 abnormal vital signs, dyspnea, pain	#2 congestive heart failure, myocardial infarction	#14 HPS simulated clinical event
	#4 complex medical-surgical scenarios		#16 a simulated clinical experience not described
	#6 fundamental level sim	#11 cardiac	
	#7 emergent patient situations	#13 pt. with chest pain	
	#10 post-op, unable to void and in pain; MVA with pain management	#15 cardiac	
	#11 respiratory	#18 cardiovascular HPS	
	12 adult emergent deteriorating patient scenario	#19 myocardial infarction	
	#13 MVA, a resp. pt., diabetes mellitus II pt.		
	#17 two patient HPS simulation complex diagnoses/ medical emergency		

Studies by ID number: 1. Aronson, Glynn, Squires (2012); **2.** Ashcraft et al. (2013); **3.** Doolen (**2012**); **4.** Gantt, (2010); **5.** Goodstone & Goodstone (2013); **6.** Haggard, (2013); **7.** Jensen, (2013); **8.** Kim & Shin, (2013); **9.** Lasater (2007); **10.** Lasater (2005); **11.** Liaw et al. (2010); **12.** Merriman et al. (2014); **13.** Meyer (2012); **14.** Mikasa, Cicero, Adamson, (2013); **15.** Nicholson (2010); **16.** Patton (2013); **17.** Radhakrishnan, Roche, Cunningham (2007); **18.** Strickland (2013); **19.** Swanson et al. (2011).

Appendix F SR Results

Table 7

CINAHL Search

#	Query	Limiters/Expanders	Last Run Via	Results
S93	S90 AND S91 AND S92	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	1,782
S92	S14 OR S15 OR S16 OR S17OR S18 OR S19 OR S20 OR S21 OR S23	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	19,036
S91	S1 OR S2 OR S3 OR S4 OR S5 OR S6 OR S7 OR S8 OR S9 OR S10 OR S11 OR S12 OR S13 OR S22	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	128,801

S90	S24 OR S25 OR S26 OR S27OR S28 OR S29 OR S30 OR S31 OR S32 OR S33 OR S34OR S35 OR S36 OR S37 OR S38 OR S39 OR S40 OR S41OR S42 OR S43 OR S44 OR S45 OR S46 OR S47 OR S48OR S49 OR S50 OR S51 OR S52 OR S53 OR S54 OR S55OR S56 OR S57 OR S58 OR S59 OR S60 OR S61 OR S62OR S63 OR S64 OR S65 OR S66 OR S67 OR S68 OR S69 OR S70 OR S71 OR S72 OR S73 OR S74 OR S75 OR S76OR S77 OR S78 OR S79 OR S80 OR S81 OR S82 OR S83OR S84 OR S85 OR S86 OR S87 OR S88 OR S89	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	299,806
S89	rubric*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	326
S88	checklist* N2 eval*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	203

S87	objective structured clinical examination	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	429
S86	OSCE	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	406
S85	MH "Competency Assessment"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	2,751
S84	MH "Psychomotor Performance"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	5,003
S83	MH "Nursing Skills"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	2,605
S82	learning N2 outcome*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	1,091
S81	instruments N2 test*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	347
S80	instrument N2 test*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	1,106

S79	instrument N2 develop*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	2,808
S78	instruments N2 develop*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	733
S77	instruments N2 construct*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	84
S76	instrument N2 construct*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	8,252
S75	competenc*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	46,612
S74	clinical N2 reason*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	1,393
S73	clinical N2 judg*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	1,543
S72	critical N2 think*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	4,214

S71	effect* N3 instruments	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	274
S70	effect* N3 instrument	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	519
S69	effect* N3 tool*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	2,181
S68	reliab*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	66,791
S67	valid*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	84,111
S66	evaluat* N3 instrument	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	2,207
S65	evaluat* N3 instruments	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	1,216
S64	outcome* N3 instruments	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	376

S63	outcome* N3 instrument	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	775
S62	measure* N3 instrument	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	3,742
S61	measure* N3 instruments	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	1,717
S60	assess* N3 instruments	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	1,658
S59	assess* N3 instrument	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	4,136
S58	assess* N3 tool*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	72,057
S57	measure* N3 tool*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	2,545
S56	outcome* N3 tool*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	823

S55	tool N3 evaluat*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	5,547
S54	observ* N2 tool*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	275
S53	teach* N3 evaluat*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	3,317
S52	teach* N3 outcome*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	328
S51	educat* N3 outcome*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	7,547
S50	MH "Clinical Competence"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	18,306
S49	MH "Decision Making, Clinical"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	15,471
S48	MH "Critical Thinking"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	3,192

S47	MH "Evaluation Research"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	17,959
S46	MH "Evaluation"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	615
S45	MH "Measurement Issues and Assessments"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	1,580
S44	MH "External Validity"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	817
S43	MH "Validity"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	6,684
S42	MH "Internal Validity"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search	762
S41	MH "Intrarater Reliability"	Expanders - Apply related words Search modes - Boolean/Phrase	Database - CINAHL with Full Text Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	4,687
S40	MH "Test-Retest Reliability"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	11,203

S39	MH "Interrater Reliability"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	16,666
S38	MH "Reliability"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	5,999
S37	MH "Instrument Construction"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	8,062
S36	MH "Instrument Validation"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	18,872
S35	MH "Content Validity"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	7,133
S34	MH "Reliability and Validity"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	9,270
S33	MH "Predictive Validity"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	2,917
S32	MH "Concurrent Validity"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	2,369

S31	MH "Criterion-Related Validity"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	3,727
S30	MH "Student Performance Appraisal"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	2,224
S29	MH "Observational Methods"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	9,185
S28	MH "Systems Validation"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	29
S27	MH "Teaching Methods, Clinical"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	1,856
S26	MH "Teaching Methods"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	18,493
S25	MH "Outcome Assessment"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	16,724
S24	MH "Outcomes of Education"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	5,836

S23	high N2 fid*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	700
S22	undergrad* N2 nurs* N2 curricu*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	240
S21	simul* N2 clinical	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	875
S20	high N2 fid* N3 mannequin*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	24
S19	high N2 fid* N3 manikin*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	22
S18	high N2 fid* N3 simulat*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	497
S17	human N3 simulat* N3 patient*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	264
S16	simulat*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	18,784

S15	MH "Computerized Clinical Simulation Testing"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	61
S14	MH "Simulations"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	5,438
S13	pre licen* N3 nurs*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	80
S12	prelicen* N3 nurs*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	93
S11	nurs* N2 educat*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	73,642
S10	undergrad* N3 nurs*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	2,706
S9	MH "Students, Nursing, Practical"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	116
S8	pre-registration N2 nurs*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	574

S7	nurs* N3 student*	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	29,399
S6	MH "Schools, Nursing"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	7,854
S5	MH "Students, Nursing, Associate"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	462
S4	MH "Students, Nursing, Diploma Programs"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	128
S3	MH "Students, Nursing, Baccalaureate"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	3,219
S2	MH "Students, Nursing"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	17,635
S1	MH "Nurses"	Expanders - Apply related words Search modes - Boolean/Phrase	Interface - EBSCOhost Research Databases Search Screen - Advanced Search Database - CINAHL with Full Text	40,992

Appendix F SR Results

Table 8

Studies Categorized by Instrument's Evaluation Method and Domain of Learning

Rubric	Checklist	Evaluation tool	Cognitive	Psychomotor
		#1 Aronson, Glynn,Squires (2012)	#1 Aronson, Glynn, Squires (2012)	#1 Aronson, Glynn, Squires (2012)
#2 Ashcraft et al. (2013)			#2 Ashcraft et al (2013)	#2 Ashcraft et al. (2013)
#3 Doolen (2012)			#3 Doolen (2012	#3 Doolen (2012
#4 Gannt (2010)			#4 Gannt (2010)	#4 Gannt (2010)
	#5 Goodstone & Goodstone (2013)		#5 Goodstone & Goodstone (2013)	#5 Goodstone & Goodstone (2013)
#6 Haggard (2013)			#6 Haggard (2013)	#6 Haggard (2013)
#7 Jensen (2013)			#7 Jensen (2013)	#7 Jensen (2013)
	#8 Kim & Shin (2013)		#8 Kim & Shin (2013)	#8 Kim & Shin (2013)
#9 Lasater (2007)			#9 Lasater (2007)	#9 Lasater (2007)
#10 Lasater (2005)			#10 Lasater (2005)	#10 Lasater(2005)
	#11 Liaw et al. (2010)		#11 Liaw et al. (2010)	#11 Liaw et al. (2010)
	#12 Merriman et al. (2014)		#12 Merriman et al.(2014)	#12 Merriman et al. (2014)
#13 Meyer (2012)			#13 Meyer (2012)	#13 Meyer (2012)
		#14 Mikasa, Cicero, Adamson, (2013)	#14 Mikasa, Cicero, Adamson, (2013)	#14 Mikasa, Cicero, Adamson, (2013)
#15 Nicholson (2010)			#15 Nicholson (2010)	#15 Nicholson (2010)
		#16 Patton (2013)	#16 Patton (2013)	#16 Patton (2013)
		#17 Radhakrishnan, Roche, Cunningham (2007)	#17 Radhakrishnan, Roche, Cunningham (2007)	#17 Radhakrishnan, Roche, Cunningham (2007)
#18 Strickland (2013)			#18 Strickland (2013)	#18 Strickland (2013)
#19 Swanson et al. (2011)			#19 Swanson et al. (2011)	#19 Swanson et al. (2011)

Appendix F SR Results

Table 9
Reliability, Validity, and Quality Assessment Data

Study ID Number Author, Year	**Statistical Analyses reported**	**Reported Reliability of instrument; Stability; Internal Consistency Reliability; Equivalence Reliability**	**Reported validity of instrument :face, content validity; Criterion Validity Construct Validity**	**Setting**	**Generalizability**	**Participant Eligibility and recruitment**
#1 Aronson, Glynn, Squires (2012)	mean SD	IRR: Phase 1 - .73, .76, .77 (3 raters) Phase 2- .83 (2 raters)	Content validity: Literature review and expert panel review	HPS sim lab in university	limited to 1 academic setting, Sr. level nursing student groups, lacks diversity in population	final yr BN
#2 Ashcraft et al. (2013)	paired sample t-tests	Internal consistency reliability: Phase I - Cronbach's α .82 for formative assessment, .91 for summative. Phase 2 - Cronbach's α .91 for formative and .92 for summative.	Content- expert panel review	school of nursing sim lab	limited due to sampling, unknown diversity of population	Sr. BN
#3 Doolen (2012)	Anovas	Stability reliability: test-retest correlation coefficient (Pearson's) ; r=.59 Internal consistency reliability: Cronbach's α =.74	None reported	nursing school sim lab	Limited to 1 site & sampling. Demographics reported with some diversity.	1st and 4th semester undergraduate BSN students
# 4 Gantt (2010)	None reported	None reported	Content- expert panel from an earlier study	sim lab	Limited to 1 population, lack of demographics and sampling	1st yr AD and Sr. BN

Study ID Number Author, Year	Statistical Analyses reported	Reported Reliability of instrument; Stability; Internal Consistency Reliability; Equivalence Reliability	Reported validity of instrument :face, content validity; Criterion Validity Construct Validity	Setting	Generalizability	Participant Eligibility and recruitment
#5 Goodstone & Goodstone (2013)	None reported	Internal consistency reliability: Cronbach's α= 0.84 Equivalence reliability: IRR rater-agreement index (RAI) .83 for the 14 student videos and .90 for sample student videos	Content validity: literature review, panel of experts Item content validity index (I-CVI) three behaviours at .75, one at .88, rest at =1.0 . Scale content validity index (S-CVI) .93	nursing sim lab	Limited due to small sample size, from same institution	PN , AD, BN with at least one semester of hospital clinical
#6 Haggard (2013)	Z-test of proportion, one sample t-test, and Pearson Correlation. Measures of central tendency, SD	Internal consistency reliability: Cronbach's α=.86-.96 from original tool / study. Cronbach's =.78- .93 current study Equivalence reliability: IRR-dependent sample t tests for each behaviour; non-significant indicating good level of agreement	States internal and content validity by establishing IRR with study instrument. Construct validity: convergent/divergent validation and factor analysis	2 university sim labs	Limited to sampling and small sample	Jr. & Sr. BN from 2 different BSN programs Excluded those under 18 and over 65

Study ID Number Author, Year	**Statistical Analyses reported**	**Reported Reliability of instrument; Stability; Internal Consistency Reliability; Equivalence Reliability**	**Reported validity of instrument :face, content validity; Criterion Validity Construct Validity**	**Setting**	**Generalizability**	**Participant Eligibility and recruitment**
#7 Jensen (2013).	Mann-Whitney U, Wilcoxon signed ranks, Spearman rho correlation.	Internal consistency reliability: Cronbach's α: LCJR scale (α =.95); for subscale: noticing (α =.88); interpreting (α =.88); responding (α =.88); and reflecting (α =.86)	content validity - none reported	university sim lab	limited to AD BN, lack of demographic data (due to limited diversity at study institution)	AD , BN
#8 Kim & Shin (2013).	χ2 test ,ANOVA. T-test, Scheffé post hoc tests	None reported	None reported	Nursing sim lab	Limited to 1 site and sampling	BN yr 2 finished a class on Women's Health Nursing & clinical training
#9 Lasater. (2007)	Mean, ANOVA	None reported	Content validity: review by experts in CJ, rubric development	nursing sim lab	limited due to sampling, no demographics, limited to 1 site	3rd term Jr.BN in medical-surgical clinical course.
#10 Lasater (2005)	ANOVA, measures of central tendency, SD	None reported	Content validity: review by expert in educ. rubric development	nursing sim lab	limited due to sampling, lack of diversity demographics, one site	Jr.BN enrolled in Nursing Care of the Acutely Ill Adult
#11 Liaw et al. (2010)	independent t-tests	None reported. No IRR-one rater	Content validity: panel of experts	nursing sim labs	Limited due to sampling	1st year BN in care of patients respiratory & cardiovascular disorders module
#12 Merriman et al. (2014)	Mann Whitney U	None reported	None reported	skills laboratory and classroom	limited to 1 site, limited due to sampling	1st year BN Adult Nursing course

Study ID Number Author, Year	Statistical Analyses reported	Reported Reliability of instrument; Stability; Internal Consistency Reliability; Equivalence Reliability	Reported validity of instrument :face, content validity; Criterion Validity Construct Validity	Setting	Generalizability	Participant Eligibility and recruitment
#13 Meyer (2012)	Dependent t tests, Kolmogorov-Smirnoff & Shapiro-Wilk statistics	None reported	None reported	university sim lab	Limited to 1 site, limited demographics, and sampling	BN in Sr. level medical-surgical class
#14 Mikasa, Cicero, Adamson, (2013)	None Reported.	Internal consistency: Cronbach's α= .97 from previous study. IRR intra-class correlation .85 Intra-rater intra-class correlation .90	content validity: review of guidelines, expert review tool development and evaluation scales	university sim lab	Limited due to sampling, one site, no demographics.	BN in adult & pediatric acute care courses
#15 Nicholson (2010)	ANOVA, measures of central tendency, SD .	Internal consistency reliability: Cronbach's α=.92; IRR .93 (pilot study); .92 (fall cohort)	Content validity- expert panel	nursing sim labs	Limited due to sampling, one site; enhanced due to random assignments of students to groups	BN in Complex Concepts of Nursing course
#16 Patton (2013)	None reported	Interrater reliability: percent agreement .85 to .89	literature review; expert panel report previous study	a simulation environment	limited due to sampling, small sample size, no demographic data	BN in clinical critical care course in the final semester
#17 Radhakrishnan, Roche, Cunningham (2007)	$\chi 2$ test	None reported	None reported	simulation lab	limited due to sampling, small sample size	Sr. BN completing a second degree

Study ID Number Author, Year	**Statistical Analyses reported**	**Reported Reliability of instrument; Stability; Internal Consistency Reliability; Equivalence Reliability**	**Reported validity of instrument :face, content validity; Criterion Validity Construct Validity**	**Setting**	**Generalizability**	**Participant Eligibility and recruitment**
#18 Strickland (2013)	t-test statistic, Pearson correlation coefficient	Stability reliability: Pearson's correlation on LCJR scores r=.314 Internal consistency reliability: Cronbach α for LCJR scale .82 IRR- one rater	None reported	Simulation Center	limited to population from 1 site, sampling	Traditional BN in adult health course Exclusion criteria: students repeating the adult health nursing course
#19 Swanson et al. (2011)	Power analysis, measures of central tendency, SD, χ2 test, ANOVA	IRR .90 & .94 rater agreement for first performance and retention performance scores respectively	content validity: literature review of guidelines	HPS simulation lab	limited due to sampling	BN in second semester nursing

Appendix F SR Results

Table 10

COSMIN Study / Instrument Quality Assessment Ratings

	BOX A	BOX B	BOX C	BOX D	BOX E	BOX F	BOX G	BOX H	BOX I	Box J
Reviewers Lisa/Barb : L/B	internal consistency	reliability	measure-ment error	content validity	Structural validity	hypotheses testing	cross-cultural validity	criterion validity	responsiveness	generalizability
study #	**L/B**	**L/B**	**L/B**	**L/B**	**L/B**	**L/B**	**L/B**	**L/B**	**L/B**	**L/B**
1	p/p	p/p	f/F	e/e						f/f
2	p/p	p/p	P/p	e/e						p/p
3	p/p	p/p	p/p	e/e						g/g
4	p/p	p/p	p/p	e/e						p/p
5	p/p	p/p	p/ p	e/e						p/p
6	p/p	p/p	p/p	e/e						f/f
7	p/p	p/p	p/p	e/e						p/p
8	p/p	p/p	p/p	f/f						f/f
9	p/p	p/p	p/p	e/e						p/p
11	p/p	p/p	p/p	e/e						p/p
12	p/p	p/p	p/p	e/e						g/g
13	p/p	p/p	p/p	e/e						p/p
14	p/p	p/p	p/p	e/e						p/p
15	p/p	p/p	p/p	e/e						g/g
16	p/p	p/p	p/p	e/e						p/p
17	p/p	p/p	p/p	p/p						f/f
18	p/p	p/p	f/f	e/e						f/f
19	p/p	p/p	f/f	e/e						f/f

Abbreviations: p= poor, f=fair, g=good, e=excellent

Studies by ID number: 1. Aronson, Glynn, Squires (2012); **2.** Ashcraft et al. (2013); **3.** Doolen (2012); **4.** Gantt, (2010); **5.** Goodstone & Goodstone (2013); **6.** Haggard, (2013); **7.** Jensen, (2013); **8.** Kim & Shin, (2013); **9.** Lasater (2007); **10.** Lasater (2005); **11.** Liaw et al. (2010); **12.** Merriman et al. (2014); **13.** Meyer (2012); **14.** Mikasa, Cicero, Adamson, (2013); **15.** Nicholson (2010); **16.** Patton (2013); **17.** Radhakrishnan, Roche, Cunningham (2007); **18.** Strickland (2013); **19.** Swanson et al. (2011)

Appendix F SR Results

Table 11

SR Search Terms Comparison Chart

CINAHL	ERIC	Medline	Embase	PsycInfo
MH “Outcomes of Education”	SU.EXACT("Outcomes of Education")		'outcome of education'/exp/mj	

MH “Outcome Assessment”	SU.EXACT("College Outcomes Assessment")
	SU.EXACT("Outcome Measures")
		Outcome* adj2 assess*	'outcome assessment'/exp/mj	outcome* N2 assess*

MH “Teaching Methods”	SU.EXACT("Teaching Methods")	*Teaching/	'teaching'/exp/mj	DE "Teaching Methods"
MH “Teaching Methods, Clinical”	SU.EXACT("Clinical Teaching (Health Professions)")	clinical adj2 teach* adj2 method*	'clinical education'/exp/mj	DE "Clinical Methods Training"
MH “Observational Methods”	SU.EXACT("Observation")	*Observation/	'observation'/exp/mj	DE "Observation Methods"
MH “Student Performance Appraisal”	SU.EXACT("Student Evaluation")	*Educational Measurement/	'education'/exp/mj	DE "Educational Measurement"
MH “Criterion-Related Validity”	(see predictive validity and validity)	*”Reproducibility of Results”/	'criterion related validity'/exp/mj	DE "Test Validity"
MH “Concurrent Validity”	(see validity)	(see above)	'concurrent validity'/exp/mj	(see above)
MH “Predictive Validity”	SU.EXACT("Predictive Validity")	(see reproducibility of results)	'predictive validity'/exp/mj	DE "Statistical Validity"

CINAHL	ERIC	Medline	Embase	PsycInfo

MH “Systems Validation”	SU.EXACT("Systems Development")
	SU.EXACT("Systems Analysis")
	system* NEAR/2 assess*
		*Systems Analysis/
		system* adj2 assess*
			'validation process'/exp/mj	DE "Systems Analysis"
				system* N2 assess*

MH “Reliability and Validity”	*(see Reliability, Validity)*	*(see reproducibility of results)*	*See other entries for reliability and validity*	*See other entries*
MH “content validity”	SU.EXACT("Content Validity")	(see reproducibility of results)	'content validity'/exp/mj	See “Test Validity”
MH “Instrument Validation”	(see textwords)	(see textword instrument terms)	'instrument validation'/exp/mj	See textwords
MH “Instrument Construction”	instrument NEAR/2 creat* instruments NEAR/2 creat*	Instrument adj2 creat* Instruments adj2 creat*	instrument near/2 creat* instruments near/2 creat*	instrument N2 creat* instruments N2 creat*
MH “Reliability”	SU.EXACT("Reliability")	See reproducibility of results	'reliability'/exp/mj	See text word
MH “Interrater Reliability”	SU.EXACT("Interrater Reliability")	Interrater adj2 reliab*	'interrater reliability'/exp/mj	DE "Interrater Reliability"
MH “Test-retest Reliability”	SU.EXACT("Test Reliability") test* NEAR/2 retest*	Test* adj2 retest*	'test retest reliability'/exp/mj	test* N2 retest*
MH “Intrarater Reliability”	intrarater NEAR/2 reliab*	intrarater adj2 reliab*	'intrarater reliability'/exp/mj	intrarater N2 reliab*
MH “Internal Validity”	SU.EXACT("Content Validity")	(see reproducibility of results)	'internal validity'/exp/mj	See “test Validity”
MH “Validity”	SU.EXACT("Validity")	See reproducibility of results	'validity'/exp/mj	See textword

CINAHL	**ERIC**	**Medline**	**Embase**	**PsycInfo**

MH “Measurement Issues and Assessments”	SU.EXACT("Measurement Techniques")
	SU.EXACT("Measurement Objectives")
		Measure* adj3 assess*
		measure* NEAR/3 issu*
			measure* near/3 assess*
			measure* NEAR/3 issu*
				measure* N2 assess*
				measure* N2 issu*

MH “Evaluation”	SU.EXACT("Evaluation")	Evaluat*	'evaluation study'/exp/mj	DE "Evaluation"
MH “Evaluation Research”	SU.EXACT("Evaluation Research")	*Evaluation Studies As Topic/	'evaluation research'/exp/mj	evaluat* N2 research*
MH “Critical Thinking”	SU.EXACT("Critical Thinking")	See textwords	'critical thinking'/exp/mj	DE "Critical Thinking"
MH “Decision Making, Clinical”	SU.EXACT("Decision Making") Clinical NEAR/3 decis*	*Decision Making/ Clinical NEAR/3 decis*	'clinical decision making'/exp/mj	DE "Decision Making"
MH “Clinical Competence”	SU.EXACT("Clinical Experience")	*Clinical Competence/	'clinical competence'/exp/mj	clinical N2 competenc*
Educat* N3 outcome*	educat* NEAR/3 outcome*	Educat* adj3 outcome*	educat* near/3 outcome*	Educat* N3 outcome
Teach* N3 outcome*	teach* NEAR/3 outcome*	Teach* adj3 outcome*	teach* NEAR/3 outcome*	Teach* N3 outcome*
Teach* N3 evaluat*	teach* NEAR/3 evaluat*	teach* adj3 evaluat*	teach* NEAR/3 evaluat*	Teach* N3 evaluat*
Observ* N2 tool*	observ* NEAR/2 tool*	observ* adj2 tool*	observ* NEAR/2 tool*	Observ* N2 tool*
Tool* N3 evaluat*	tool* NEAR/3 evaluat*	tool* adj3 evaluat*	tool* near/3 evaluat*	Tool* N3 evaluat*
Outcome* N3 tool*	outcome* NEAR/3 tool*	outcome* adj3 tool*	outcome* NEAR/3 tool*	Outcome* N3 tool*

CINAHL	ERIC	Medline	Embase	PsycInfo
Measure* N3 tool*	measure* NEAR/3 tool*	measure* adj3 tool*	measure* NEAR/3 tool*	Measure* N3 tool*
Assess* N3 tool*	assess* NEAR/3 tool*	assess* adj3 tool*	assess* NEAR/3 tool*	Assess* N3 tool*
Assess* N3 instrument	assess* NEAR/3 instrument	assess* adj3 instrument	assess* NEAR/3 instrument	Assess* N3 instrument
Assess* N3 instruments	assess* NEAR/3 instruments	assess* adj3 instruments	assess* NEAR/3 instruments	Assess* N3 instruments
Measure* N3 instruments	measure* NEAR/3 instruments	measure* adj3 instruments	measure* NEAR/3 instruments	Measure* N3 instruments
Measure* N3 instrument	measure* NEAR/3 instrument	measure* adj3 instruments	measure* NEAR/3 instrument	Measure* N3 instrument
Outcome* N3 instrument	outcome* NEAR/3 instrument	outcome* adj3 instrument	outcome* NEAR/3 instrument	Outcome* N3 instrument
Outcome* N3 instruments	outcome* NEAR/3 instruments	outcome* adj3 instruments	outcome* NEAR/3 instruments	Outcome* N3 instruments
Evaluat* N3 instrument	Evaluat* NEAR/3 instrument	Evaluat* adj3 instrument	Evaluat* NEAR/3 instrument	Evaluat* N3 instrument
Evaluat* N3 instruments	Evaluat* NEAR/3 instruments	Evaluat* adj3 instruments	Evaluat* NEAR/3 instruments	Evaluat* N3 instruments
Valid*	Valid*	Valid*	Valid*	Valid*
Reliab*	Reliab*	Reliab*	Reliab*	Reliab*
Effect* N3 tool*	effect* NEAR/3 tool*	effect* adj3 tool*	effect* NEAR/3 tool*	Effect* N3 tool*
Effect* N3* instrument	effect* NEAR/3 instrument	effect* adj3 instrument	effect* NEAR/3 instrument	Effect* N3* instrument

CINAHL	**ERIC**	**Medline**	**Embase**	**PsycInfo**
Effect* N3 instruments	effect* NEAR/3 instruments	effect* adj3 instruments	effect* NEAR/3 instruments	Effect* N3* instruments
Critical N2 think*	critical NEAR/2 think*	Critical adj2 think*	critical NEAR/2 think*	Critical N2 think*
Clinical N2 judg*	critical NEAR/2 judg*	critical adj2 judg*	critical NEAR/2 judg*	Clinical N2 judg*
Clinical N2 reason*	critical NEAR/2 reason*	critical adj2 reason*	critical NEAR/2 reason*	Clinical N2 reason*
Competenc*	Competenc*	Competenc*	competenc*	Competenc*
Instrument N2 construct*	instrument NEAR/2 construct*	instrument adj2 construct*	instrument NEAR/2 construct*	Instrument N2 construct*
Instruments N2 construct*	instruments NEAR/2 construct*	instruments adj2 construct*	instruments NEAR/2 construct*	Instruments N2 construct*
Instruments N2 develop*	instruments NEAR/2 develop*	instruments adj2 develop*	instruments NEAR/2 develop*	Instruments N2 develop*
Instrument N2 develop*	instrument NEAR/2 develop*	instrument adj2 develop*	instrument NEAR/2 develop*	Instrument N2 develop*
Instruments N2 test*	instruments NEAR/2 test*	instruments adj2 test*	instruments NEAR/2 test*	Instruments N2 test*
Instrument N2 test*	instrument NEAR/2 test*	instrument adj2 test*	instrument NEAR/2 test*	Instrument N2 test*
Learning N2 outcome*	learning NEAR/2 outcome*	learning adj2 outcome*	learning NEAR/2 outcome*	Learning N2 outcome*
MH “Nursing Skills”	nurs* NEAR/2 skill*	nurs* adj2 skill*	nurs* NEAR/2 skill*	nurs* N2 skill*

CINAHL	**ERIC**	**Medline**	**Embase**	**PsycInfo**
MH “Psychomotor Peformance”	SU.EXACT("Psychomotor Skills")	*Psychomotor Peformance/	'psychomotor performance'/exp/mj	DE "Perceptual Motor Processes"
MH “Competency Assessment”	See competenc*	See competenc*	See competenc*	DE "Minimum Competency Tests"
Rubric*	rubric*	Rubric*	rubric*	Rubric*
Checklist* N2 eval*	checklist* NEAR/2 eval*	checklist* adj2 eval*	checklist* NEAR/2 eval*	Checklist* N2 eval*
OSCE	osce	Osce	osce	osce
Objective structured clinical exam	Objective structured clinical exam	Objective structured clinical exam	Objective structured clinical exam	objective structured clinical exam

Appendix F SR Results

Table 12

Studies Categorized by Research Design

Experimental randomized -controlled trial Studies	**Quasi-experimental**: -non-randomized controlled trial - comparative design before and after interrupted time series	**Pilot study** : -experimental -quasi-experimental	**Non-experimental** : - pre and post cohort -instrument development -descriptive
12	**8** pretest-posttest design	**4** quasi-experimental pilot study	**1** instrument development study
15 posttest-only design	**11** cross-over intervention posttest design	**5** quasi-experimental pilot study posttest only	**2** descriptive study
18 pretest/post-test design	**13** non-randomized controlled trial	**17** quasi-experimental pilot study 2 group posttest	**3** methodological instrument development
19 posttest only design			**6** correlational design
			7 descriptive study
			9 exploratory mixed methods- a qualitative-quantitative-qualitative design
			10 exploratory qualitative and quantitative methods
			14 instrument development study
			16 descriptive study

Studies by ID number: 1. Aronson, Glynn, Squires (2012); **2.** Ashcraft et al. (2013); **3.** Doolen (**2012**); **4.** Gantt, (2010); **5.** Goodstone & Goodstone (2013); **6.** Haggard, (2013); **7.** Jensen, (2013); **8.** Kim & Shin, (2013); **9.** Lasater (2007); **10.** Lasater (2005); **11.** Liaw et al. (2010); **12.** Merriman et al. (2014); **13.** Meyer (2012); **14.** Mikasa, Cicero, Adamson, (2013); **15.** Nicholson (2010); **16.** Patton (2013); **17.** Radhakrishnan, Roche, Cunningham (2007); **18.** Strickland (2013); **19.** Swanson et al. (2011).