Running head: CLINICAL JUDGMENT CONSTRUCT VALIDITY IN SIMULATION

An examination of the construct validity of a clinical judgment evaluation tool in the setting of high-fidelity simulation

By

Stephanie Sideras

A Dissertation

Presented to
Oregon Health & Science University
School of Nursing
in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

December 20, 2007

APPROVED:

Christine A. Tanner, RN, PhD, Professor, Dissertation Chair

Barbara Gaines, RN, EdD, Professor, Committee Member

Nancy Perrin, PhD, Professor, Committee Member

Dannielle Stevens, PhD, Professor, Portland State University, Committee Member

Saundra Theis, RN, PhD, Interim Dean, School of Nursing

ACKNOWLEDGMENTS

I would like to gratefully acknowledge the hard work, support and patience of my dissertation committee. To my dissertation chair, Dr. Christine A. Tanner, my thanks for providing thought provoking and thought clarifying messages needed to help me through the maze of ideas this project created. Your model of clinical judgment was my starting point and my map and I profoundly appreciate being given the opportunity to research clinical judgment under your guidance. I also truly appreciate your manner of balancing encouragement and deadlines to keep my feet walking forward to the heart of the matter. To Dr. Barbara Gaines, my thanks for helping conceptualize a dissertation as an achievable first piece of research while at the same time applying your experience with education, nursing and clinical judgment to maintain the rigor of the work. To Dr. Nancy Perrin, my thanks for demonstrating true educational scaffolding, your patience while teaching me statistical methods, and your patience during those subsequent afternoons in your office is a role I strive to emulate. I appreciate your awareness of both the power of self-discovery and the power of reassurance. To Dr. Dannielle Stevens, my thanks for helping me understand the nature of rubrics, both from the view of the educator and the view of the learner. Transparency of expectations is a powerful educational concept and one that I will continue to explore. I also want to extend my thanks to Dr. Kathie Lasater, the author of the clinical judgment rubric. The clarity of Dr. Lasater's prior research was an invaluable aid and her support and encouragement helped me understand the nature of collaboration. Finally, I wish to thank my mentor, Dr. Saundra Theis, for knowing how well I would fit in the world of high fidelity simulation, which is the aspect of nursing education that has become my passion.

DEDICATION

To my family -

My parents, Arthur T. and Joyce C. Sideras

My brother and his family - Scot, his wife Leah, and sons Sam and Tom

The sister of my heart and her family – Suzy, her husband Kevin and daughter Kelly

To my students –

All of whom make every day an exciting learning opportunity

ABSTRACT

Title: An examination of the construct validity of a clinical judgment evaluation tool in the setting of high-fidelity simulation

Author: Stephanie Sideras

APPROVED: ____

Christine A. Tanner, RN, PhD, Dissertation Chair

The purpose of this study was to examine the construct validity of the Lasater Clinical Judgment Rubric (LCJR) as a method of evaluating clinical judgment. The primary aims of the study were to examine the empirical evidence for two testable theoretical relationships present in the Tanner research-based clinical judgment model. A secondary aspect of the study was the exploration of clinical judgment assessments made by faculty in comparison to those made by students. A baccalaureate nursing program requiring two years of upper division course work provided the convenience sample. Known-groups methodology was used to compare the clinical judgment performances of graduating seniors (N = 24) to those of end-of-year juniors (N = 22). The setting was a high-fidelity simulation lab with each student participating in three scenarios of increasing complexity. The active component of each simulation and the debriefing was recorded. Nursing faculty raters were blind to the educational status of the student. The first hypothesis, that graduating seniors would demonstrate a higher level of clinical judgment than end-of-year juniors was supported. With performance averaged across all three of the simulation scenarios, the z-score probability interpretation of the effect size difference between these two groups found that 81% of the juniors demonstrated a level of clinical judgment that was less than the average of the senior group. While this study provides strong support for the construct validity of the LCJR, further research needs to be directed toward methods of improving inter-rater agreement, which was modest in this study. The second hypothesis was that as simulation complexity increased, student performance in clinical judgment, as measured on the LCJR, would decrease. This hypothesis was examined through repeated measures ANOVA, and was not supported. The possible explanations for the lack of interaction between complexity and clinical judgment include possible error in the measurement or implementation of complexity, a practice effect as the students went directly from one simulation to the next, or insufficient power in the study design. Future research in this area needs to address our understanding of situational complexity. The third hypothesis examined the relationship between student self-evaluation of clinical judgment and faculty rating. A Pearson r correlation demonstrated significance only in the dimension of noticing. While paired t-test analysis found significant differences between faculty and both groups of students, less difference was present between faculty and senior students. An analysis of variance determined student grouping did not significantly impact the accuracy of self-assessment. Future research should be directed toward improving student awareness of the expected the standard of performance. Conclusion: Preliminary evidence supports the construct validity of the Lasater Clinical Judgment Rubric.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

As a practice profession, the discipline of nursing has always required that nurses be both competent in their knowledge of nursing science as well as competent in their ability to apply that knowledge base in a safe and effective manner. As the process of clinical judgment describes how a thinking nurse responds to the specific clinical challenges encountered when providing nursing care to patients, the construct of clinical judgment embodies the core of the profession. Development of competent clinical judgment within nursing students is thus an essential goal in nursing education. However, the ability to measure competence in clinical practice has been an elusive goal for nursing educators (Watson, Stimpson, Toping & Porock, 2002; Wooley, 1977). As the profession entered the 21st century, new challenges have emerged that made the goal of competence assessment both more complex and more imperative. These challenges have come from the practice environment of nursing, from the educational discipline and from within the profession itself.

From the practice environment the challenge to nurses is to respond to the explosion of healthcare knowledge and technology. Patient treatments and the adjunct technology used during care has become more complex. As the complexity and acuity of the in-hospital patients increased, thus requiring more services, the economic response is to decrease the length of stay. The presence of the RN was a key factor identified in maintaining patient safety in a practice environment necessitating sicker patients be treated faster (Aiken, Clarke, Sloane, Sochalski, & Silber, 2002). The challenge to nurse educators is to train nurses capable of competent clinical judgment in this type of demanding practice environment.

From the educational arena, the challenge to nurse educators is to respond to the explosion of knowledge regarding how people learn. The educational paradigm is shifting from a focus on teaching to the prioritization of what students are learning (Barr & Tagg, 1995). Nursing faculty are required to design learning activities and develop learning environments that actively engage students in the process of constructing their own understanding. Valid outcome measures are a necessary component in the learning paradigm in order to determine if the student has acquired a sufficient level of understanding. In today's complex healthcare environment, nurse educators need strong outcome measures that effectively assess a student's abilities to respond to the demands of a clinical situation. Clinical judgment as a reflection of both how a nurse thinks and how a nurse acts has the potential to be a robust indicator of competence in practice.

The expanding nursing shortage is the challenge that emerged from within the nursing profession. Nurse educators are being asked to respond to the challenges from the practice environment and from the educational arena by developing new curricula that address increasingly complex nursing knowledge in a manner that acknowledges the reciprocal nature of the teaching learning process and accomplish this task in the setting of a nursing shortage. It is the incipient shortage that adds an overlay of urgency to this process. Every graduate of every nursing program across the country has the potential for making a significant difference to the populations they serve. The hazard of inappropriately failing a student will have an impact within communities. But the hazard of inappropriately passing a student who is incapable of competent clinical judgment has not diminished; nurse educators continue to have a responsibility to the public. Accurate assessment of clinical competence becomes more imperative in an educational setting where nursing faculty are being asked to produce more graduates, faster, that are competent to function in an increasingly complex healthcare environment.

Background of the Study

In 1988, then Secretary of Education William Bennett issued an executive order "that institutions or programs confer degrees only on those students who have demonstrated educational achievement as assessed and documented through appropriate measures" (US Department of Health as cited in Banta, 2001). As a result of this executive order, accrediting organizations altered their review criteria to focus on student learning as a program outcome. One of the earliest and most publicly acknowledged outcome measures identified was critical thinking. The National League for Nursing Accrediting Commission established critical thinking as one of their core performance indicators of program achievement in 1992 (NLNAC, 1992). The accrediting arm of the American Association of Colleges of Nursing (AACN) followed suit in 1995. Currently, the Oregon Nurse Practice Act continues to require critical thinking as a component in the curriculum of a registered nurse.

However, the early identification of critical thinking as a programmatic outcome measure became problematic for the nursing profession. Although the American Philosophical Association had published a cross-disciplinary definition of critical thinking in 1990, a subsequent survey of 55 schools of nursing found 10 different definitions in use (Videbeck, 1997). Throughout the 1990's, nurse educators were found to most commonly associate the definition of critical thinking with scientific problem solving (Gordon, 2000; O'Sullivan, Blevins-Stephens, Smith & Vaughan-Wrobal, 1997), with some educators viewing it as a purely cognitive process and many others viewing it as a process containing both cognitive and affective components (Videbeck, 1977). However, from its inception as a programmatic outcome measure, critical thinking assessments of nursing students have demonstrated an inconsistent relationship between nursing education and critical thinking ability (Brunt, 2005; Staib, 2003). While many researchers attributed the inconsistencies to the absence of a nursing specific definition of critical thinking (Gordon, 2000; Scheffer & Rubenfeld, 2000), other researchers identified the rational-linear approach itself as limiting and insufficient to explain nursing practice (Jones & Brown, 1991). The absence of any nursing specific instruments to measure critical thinking following the publication of the nursing specific definition of critical thinking developed by Scheffer and Rubenfeld in 2000, lends support to the position that critical thinking is an inadequate construct to explain nursing practice.

Clinical judgment is "an interpretation or conclusion about a patient's needs, concerns or health problems and/or the decision to take action (or not), and to use or modify standard approaches, or to improvise new ones as deemed appropriate by the patient's response" (Tanner, 1998, p. 19). This definition of clinical judgment sets it apart from critical thinking and identifies it as case-based, contextually bound, interpretive reasoning (Tanner, 2007). Clinical judgment is an integrative way of knowing. It is integrative because the nurse uses multiple types of knowledge – theoretical, practical, interpersonal, experiential – within the frame of reference established by a specific patient. The development and evaluation of clinical judgment is the focus of this research.

The research-based model of clinical judgment developed by Tanner (2000) describes the process used by each individual nurse to best respond to situations encountered in practice. The model was the result of a synthesis of 191 research studies (Tanner, 2006) that examined three broad areas: 1) the reasoning patterns used by nurses 2) the roles of knowledge and experience on reasoning patterns 3) other factors that affect clinical reasoning patterns. Tanner identified five consistent conclusions in this synthesis of literature and used these to develop a model that explains how nurses think in the rapidly changing environment of clinical practice.

The process of clinical judgment is influenced by context and complexity from two major sources, the patient and the nurse. Both patient and nurse bring their lived experiences with health alterations to the encounter. This context, which is inclusive of background knowledge and practical experience, strongly influences the clinical judgment process. The complexity of the interactions between the background and context impact the nurse's initial grasp of the situation, and this initial grasp is what initiates the clinical judgment process by directing what the nurse notices. Interpretation of the data noticed follows next, and then the nurse responds, or acts, on the information. The final dimension of the process is reflection, which enables the link between the current clinical situation and the nurse's past, as well as future, experiences with similar situations.

In 2005 a rubric was created (Lasater) to facilitate the evaluation of clinical judgment. The content of the rubric emerged in part from the theoretical relationships present in the Tanner model and in part from Lasater's observations of the responses of students to clinical challenges encountered in a high-fidelity simulation setting. The

Lasater Clinical Judgment Rubric uses the dimensions of clinical judgment – noticing, interpreting, responding and reflecting, to provide structure for the eleven separate indicators necessary to fully describe the process. The rubric also differentiates four levels of ability – beginning, developing, accomplished and exemplary. The value of this rubric lies in the fact that it translates a research-based process that describes how nurses function in clinical practice into the format of an outcome measure. Lasater's 2005 work focused on the development and pilot testing of the clinical judgment rubric. This study addresses an examination of the construct validity.

The Importance of Clinical Judgment as an Outcome Measure

The construct of clinical judgment emerged from a desire to understand what it is to "think like a nurse". The value of clinical judgment as an outcome measure rests in the fact that it is an integrative construct. Clinical judgment assessment examines not only how the nurse thinks but also how the nurse responds to the clinical practice situation and the nurse's willingness to reflect and learn from the encounter. Thus, clinical judgment assessment would provide deep insight into the various knowledge bases of nurses, and their ability to apply that understanding through clinical actions.

Clinical judgment evaluation is critical to future nursing education research, since it provides an opportunity for comparison among levels of practice. The qualitative work of Benner (1984, 2004) demonstrates that the acquisition of expertise in clinical judgment is developmental. The clinical judgment of novice nurses is dramatically different from that of experts. There are stages of development identified as occurring during the educational process as well as those that benchmark the ongoing career development of the nurse. The clinical judgment rubric is designed to differentiate the varying levels of performance.

While competency assessment will always remain an essential responsibility of nurse educators, the clinical judgment rubric facilitates the inclusion of the learner in the evaluation process. Learning is enhanced when the learner understands the goal, or outcome of the activities (Bransford et al., 2000). The transparency of the standards and expectations for performance delineated in the rubric become helpful to faculty and students as both parties reflect on the student's current level of performance and identify priority areas for growth in clinical judgment ability. Deliberate practice facilitates the acquisition of expertise (Ericsson, 2004). Deliberate practice is a process of avoiding stagnation in learning, a stagnation that reinforces that the current 'good enough' is sufficient. When used by nursing faculty, deliberate practice is a method of scaffolding ongoing challenges to engage learners in the process of improvement. Deliberate practice coupled with feedback, fosters the development of expertise.   Nurse educators can use clinical judgment performance assessment to direct and focus the learning activities of the student using deliberate practice to advance student ability.

Simulation as a Setting for Clinical Judgment Outcomes Assessment

High-fidelity simulation is an optimal setting for the evaluation of student clinical judgment performance. Simulation is an educational method where aspects of a real-world domain are artificially abstracted and replicated in order to safely achieve a defined learning goal (Gaba, 2004; Hertel & Millis, 2002). The level of fidelity refers to how closely the simulation replicates reality. Fidelity is determined by the number of elements of reality that are replicated as well as the amount of error allowed between each element in the simulation and the real world (Gaba, 2004). High-fidelity simulation typically uses a manikin that closely replicates a patient; heart, lung and bowel sounds are audible, peripheral pulses are palpable, spontaneous vocalizations are possible and realistic therapeutic modalities can be instituted, such as cardiac monitoring or intravenous fluid therapy. Further, the setting of a high-fidelity simulation can be altered to reflect any practice environment appropriate to nursing. Common settings developed for simulation practice include a hospital room, an operating room, an obstetric delivery room, and a home living room.

According to LeBreck (1989) the examination of clinical judgment "must include collection and assessment of data by the decision maker at the speed and level of uncertainty found in actual practice, the selection and performance of therapeutic actions, reactions to changes in the patient resulting from therapy and ongoing patient management. " (p. 43). The realism present in a high-fidelity simulation setting promotes this type of evaluation. The level of uncertainty can be established by the evaluator, as well as the types of problems encountered by the learner. The ability to observe students making choices and monitoring the ongoing status of the patient is possible in the high-fidelity setting. The student learner enters the setting as the fully responsible registered nurse. The clinical problem presented in the simulation is the responsibility of the nurse to resolve. The interaction between the nurse and the patient is the central focus. Nursing faculty receive the benefit of direct observation of the learner without the impediment of their presence in the room as a distraction from the work of the student with the patient. The use of a patient manikin resolves the safety issues that historically precluded this level of independent practice by the student.

It is this level of independence in active practice and direct observation that establishes high-fidelity simulation as a method apart from prior types of clinical judgment assessment. Historically, two assessments methods were used in clinical judgment evaluation. One type used paper-and-pencil branched simulations that provided the learner with an initial situation and required choices be made regarding what data to collect and actions to implement. A second type required the nurse to ‘think aloud’ either during the situation or retrospectively and these verbalizations were subsequently analyzed. Both methods were found to be inadequate in responding to and measuring the contextualized knowledge that differentiates levels of clinical judgment (Holzemer, 1986; LeBreck, 1989; Tanner, 1987).

By contrast, the active process of high-fidelity simulation inherently reveals the contextual knowledge of the participants. The fidelity of the clinical problems presented is limited only by the level of technology available in the manikin, the clinical expertise of the designer of the simulation and the imagination of the props engineer. In a high-fidelity clinical practice session the setting is controlled, and the educational background of the participants is known. The educator, using the vehicle of the students’ preparatory information, structures the expectations that frame the nurses’ initial grasp. The students’ ability to respond to the clinical situation and the process of their response reveals their level of clinical judgment. The underlying assumption is that students are demonstrating the same level of clinical judgment in their interactions in the simulation theater that they would during their interactions with patients at the clinical setting.

There are several other logistical advantages of high-fidelity simulation as a location for clinical judgment performance assessment. In clinical practice, patient availability is limited and the state of patient health is unpredictable. In the simulation setting, nursing faculty can both identify the common clinical competencies essential for all students and schedule the practice and assessment of these events. Also, the control possible in a simulation theater allows for the delivery of equal testing situations for all students, which is not possible in clinical practice. Situational equality in clinical performance assessment is important as it helps faculty to maintain a stable level of competency expectations.

Importance of Construct Validation in Clinical Judgment

As the construct of clinical judgment represents the essential core of how nurses think and act, the validation of an evaluation tool designed to measure clinical judgment is important. Faculty have a responsibility to assess student ability consistently and accurately. Fulfillment of the responsibility of student assessment cannot be honestly discharged without a research base substantiating the validity of the evaluation instrument. This study represents a priority step in the process of developing a valid method of assessing a core clinical competency in nursing.

Construct validation refers to the process of examining the degree to which evidence and theory support the interpretations of test scores within a specific test setting (*Standards*, 1999). There are three distinct steps in the process of construct validation (Carmines & Zeller, 1979). The first step is to link the theoretical relationships between the concepts. The second step is to examine the empirical relationships between the measures of the construct. The third step is the interpretation of the empirical evidence as it relates back to the theoretical concepts. The construct validity argument developed within this study used this framework in the assessment of the evidence. The theoretical relationships present in the research-based clinical judgment model offer two testable assumptions important to a construct validation study:

1) Nurses with greater practical experience and greater levels of nursing specific knowledge will demonstrate higher levels of clinical judgment.
2) Level of clinical judgment ability demonstrated will be related to the level of situational complexity present.

Known groups approach was the method of construct validation testing used to reveal the empirical relationships between these concepts. In this study, the two groups compared on level of clinical judgment ability were end-of-year junior nursing students and graduating seniors. If the test results from the Lasater Clinical Judgment Rubric were to find the differences known to exist between these groups, this then becomes a source of evidence in support of the construct validity of the tool.

The Problem

Facilitating the development of nurses who are competent in clinical judgment is the primary goal of nursing education programs. Assessing the competence of the clinical judgments that nursing students make is the primary responsibility of nursing faculty. A competence assessment is an inherent component of the contract held between the profession of nursing and the general public (Boland & Laidig, 2001) and it is required by nursing's accrediting and licensing agencies (AACN, 1998; NLNAC, 1992; OSBN, 2001). However, competence in the active practice of nursing as demonstrated by the level of clinical judgment displayed by students has been historically limited by our misconceptions of the concept and the setting available for performance appraisals. The recent work of Lasater (dissertation, 2005) in developing an instrument to evaluate a learner's stage of clinical judgment shows promise but the tool is without established reliability or validity.

Purpose of the Study

The purpose of this study was to examine the construct validity of the Lasater Clinical Judgment Rubric (LCJR) as a method of evaluating clinical judgment. The primary aims of the study were designed to facilitate the examination of the empirical evidence for two testable theoretical relationships present in the research-based clinical judgment model. The first hypothesis tests the assumption that level of knowledge and practical experience influences performance of clinical judgment.

1. Graduating seniors will demonstrate a higher level of clinical judgment than end-of-year juniors as rated by faculty and measured on the Lasater Clinical Judgment Rubric.

The theoretical relationships between the dimensions of clinical judgment present in the model support this hypothesis; senior students should have a greater nursing specific knowledge base and have had the opportunity to acquire more practical experience then junior nursing students. Background research also supports the conclusion that the process of nursing education positively affects the development of clinical judgment (Adams, 1999; Kintgen-Andrews, 1991; Staib, 2003).

The second hypothesis of this study explored the interaction between situational complexity, level of nursing expertise and clinical judgment performance. The research-based model of clinical judgment suggests that the nurse's initial grasp of the issue at hand is influenced by the situational context, the nurse's background and the relationship with the patient. Increasing the situational complexity should influence the demonstration of nursing clinical judgment. Even expert nurses, when placed in a new environment and presented with highly complex situations, are not likely to demonstrate the same level of clinical judgment they exhibit when faced with complex situations in a familiar environment. The possibility of a relationship between complexity and level of nursing expertise on clinical judgment ability was explored through the following hypothesis:

2. As simulation complexity increases, student performance in clinical judgment, as measured on the Lasater Clinical Judgment Rubric, will decrease.

All students are expected to show a decrease in level of clinical judgment performance as situational complexity increases. However, senior nursing students, with their greater level of domain specific nursing knowledge, are expected to maintain higher levels of clinical judgment performance than junior students when faced with situations of increasing complexity. If this pattern of response to complexity were found, it would support the theoretical foundation present in the clinical judgment model. It would also support the use of the high-fidelity simulation setting as a location for performance assessment with the inherent ability to control aspects of complexity.

A secondary aspect of the study was the exploration of the clinical judgment assessments made by faculty in comparison to those made by students. In a learner-centered paradigm, outcomes evaluation does not remain solely the prerogative of faculty. Learners are involved in the process and, in fact, a priority outcome in learner-centered education is development of the ability to reliably and validly self-evaluate (Weimer, 2002). This comparison was examined through the following hypothesis:

3. There is a relationship between student self-evaluation of clinical judgment and faculty rating of that student's level of clinical judgment.

If faculty ratings of student clinical judgment were similar to student self-evaluation of level of ability, this would indicate both groups had a similar understanding of the rubric. This would lend support to validity as it supports the clarity of the rubric descriptors to both faculty and students. Discrepancies between faculty ratings and student ratings would suggest either further work addressing clarity of the standard of performance on the part of the faculty or further work addressing the clarity of self-evaluation on the part of the student, or further work required by both groups.

In the next chapter, the review of literature presented examines the theoretical foundations used in this study. The types of evidence necessary to support construct validity are reviewed. Four models of clinical judgment are examined, along with the historical methods used to measure clinical judgment ability, to support the choice of using the Tanner (2006) model and the Lasater (2005) rubric. The theoretical relationships between the dimensions of noticing, interpreting, responding and reflecting that are present in current research are examined in depth. Finally, the literature related to learning in high-fidelity simulation is reviewed to establish this setting as the optimal location for this study.

CHAPTER TWO

REVIEW OF LITERATURE

When applied to psychometrics, the term validity refers to how well a research instrument measures the concept it is designed to measure (Nunnally & Bernstein, 1994). The American Educational Research Association (AERA), the American Psychological Association (APA) and the National Council on Measurement in Education (NCME) in the *Standards for Educational and Psychological Testing* define validity as "the degree to which evidence and theory support the interpretation entailed by proposed use of tests" (1999, p. 9). Historically, validity has been subdivided into three different types, content, criterion-related and construct (Nunnally & Bernstein, 1994). The 1999 *Standards* rejects this division and states clearly "validity is a unitary concept" (p. 11) and while there are different types of validity evidence, there are not different types of validity. The five different types of validity evidence all have equal value in the examination of the validity of the construct. Validity as a unitary concept is now construct validity and the accuracy of the construct is evaluated through examination of the different types of evidence.

The process of validation, as specified in the *Standards* (1999), is the development of an evidence base that logically supports the proposed score interpretation. The purpose of the review of literature is to present the evidence base that will be used in the examination of the construct validity of the clinical judgment rubric in the setting of high-fidelity simulation. The first section contains a discussion of construct validity, the five types of validity evidence and common threats to validity. Section two presents the theoretical foundations of clinical judgment as a construct and presents the evidence supporting the relationships between the concepts. A conceptual model for learning in the setting of high-fidelity simulation is presented in section three and through a discussion of the component parts of the conceptual model will address how active practice in the HFS setting facilitates development of clinical judgment. Finally, section four summarizes the literature related to learning in high fidelity simulation, identifies the gaps in knowledge as well as the advantages of high-fidelity simulation as a location for the assessment of clinical judgment.

Construct Validity

Psychometric assessment of construct validity has become increasingly rigorous over the past 50 years. Goodwin (1997) summarized an overview of the historical changes in the conceptualization of validity. In 1946 Guilford stated "in a very general sense, a test is valid for anything with which it correlates" (p429) which reflected the position that validity was information that related a test to its specific aim. Cronbach and Meehl (1955) modified this position by dividing validity into three categorical sub-types: content, criterion-related and construct, with predictive and concurrent validity identified as subtypes of criterion-related validity. Campbell and Fiske (1959) extended this work by developing definitions for convergent validity and discriminant validity as types of construct validity.

While these categorizations of validity types are still current in many texts (Nunnally & Bernstein, 1994), during the 1980's this categorical approach to validity began to change. Researchers began to discuss the limitations of such a categorical approach to validation suggesting that it tended to reduce validity work to a checklist approach (Goodwin, 2002). The compartmentalization of validity fostered the misconception that validity was the property of the measure rather than a property of the scores that result from the application of a measure for a specific purpose and with a specific population. Messick (1988) revised the definition of validity as follows: "Validity is an overall evaluative judgment, founded on empirical evidence and theoretical rationales, of the adequacy and appropriateness of inferences and actions based on test scores" (p33). While the 1985 edition of the *Standards* continued to use category labels, it did state that the presence of such labels should not imply distinct types of validity, and as noted above, in the 1999 edition of the *Standards*, the categorical division of validity is absent, and validity is discussed as the unitary concept of construct validity.

A validity argument as discussed by Downing (2003) necessitates presentation of a chain of evidence that clearly links theory, hypotheses and logic in ways that either support or refute the reasonableness of the desired interpretations. The 1999 *Standards* identified five types of validity evidence:

1. Evidence based on test content
2. Evidence based on response processes
3. Evidence based on internal structure
4. Evidence based on relations to other variables
5. Evidence based on consequences of testing

While each type of evidence is equally valuable, the design of the instrument or the design of the study may make some types of evidence more critical than others.

Evidence based on test content examines the relationship between the content on the instrument and the construct it is intended to measure (*Standards*, 1999). This type of evidence can be obtained from a logical analysis of the content domain as represented in the literature when compared to the content represented on the instrument. It can also be elicited from a panel of experts who evaluate the sufficiency, relevancy and clarity of the components included in the instruments (Goodwin, 2002). Appropriate depth and breath of evidence are important whether it reflects the scope of the literature review or the panel of assessing experts (Polit & Beck, 2006; *Standards*, 1999; Stewart, Lynn & Mishel, 2005).

Evidence based on response processes concerns the fit between the construct being tested and the nature of the performance or response the examinee actually engages in (*Standards*, 1999). Evidence of this type determines whether examinees are truly thinking and responding to the test material or rather following an established algorithm. Sources of this type of evidence are usually the examinees, either through verbal investigations of their performance strategies or maintaining records that document the development of a response (Goodwin, 2002). Evidence relating to response process is not limited to a focus on the examinee. It is also vital to examine the responses processes of observers/raters to determine if they are applying criteria as intended and not responding to irrelevant or extraneous factors that are not part of the planned interpretation of the scores.

Evidence based on internal structure addresses the degree to which the relationships as operationally defined on the instrument components match the construct on which they were based (*Standards*, 1999). The question asked here is whether the components on the instrument are homogenous but distinct aspects of the construct or if there is too much overlap within a dimension. Most typically this type of evidence is gathered through factor analysis but depending upon the instrument and how the test will be applied, an analysis of response patterns or differential item analysis may be sources of evidence.

Evidence based on relationships to other variables addresses questions concerning the nature and extent of the relationship between instrument scores and other variables. Depending on the other variable type, this relationship could reflect a convergent, concurrent, discriminant or predictive relationship (*Standards*, 1999). Known group comparison studies are an example of this type of evidence. Known group comparisons use a hypothesis testing approach. The validity of the measure is revealed by examining the difference in performance by groups known to have different levels of the construct of interest. If the scores on the measure are high in groups known to have high levels of the construct and low in groups known to have low levels of the construct, the instrument has some evidence of validity (Waltz, Strickland & Lenz, 2005).

The final type of validity evidence is that based on consequences of testing, which examines the impact on the examinee as a result of the assessment score. Messick (1994) incorporated consequences into measurement validity in order to examine the possibility of intended positive effects as well as unintended negative effects with performance assessment. This type of validity evidence is controversial according to Goodwin (2002) because the evidence goes beyond psychometric boundaries and into policy decisions. The 1999 *Standards* is careful to address this exact issue stating that evidence about consequences can inform validity decisions but care must be taken to distinguish between evidence that is directly related to validity and evidence which falls into the realm of policy.

According to Downing and Haladyna (2004) any factor that can interfere with the meaningful interpretation of data is a threat to validity. However, the global nature of this viewpoint increases the difficulty of a threat analysis. Messick (1994) identified two major sources of validity threats: construct under-representation and construct-irrelevant variance. Construct under-representation describes the degree to which an instrument fails to capture important aspects of the construct (*Standards*, 1999). Construct-irrelevant variance describes the degree to which instrument scores are impacted by processes extraneous to the intended construct (*Standards*, 1999). This study will analyze validity threats using Messick's typology.

This summary of construct validity literature presents the current view of psychometrics that holds validity as a unitary concept and that concept is construct validity. The five different types of validity evidence are defined and discussed because it is through the support of the evidence that the strength of any validity argument is established. It is important to note here that validity cannot be proven, and it cannot be established through any one study, it is an ongoing process. Validity does not lie in the measure of the construct but rather in the interpretation of the data that emerges from the application of a specific instrument designed to assess a specific construct in a specific population.

The Theoretical Foundations of Clinical Judgment

The theoretical foundations of clinical judgment should be clearly visible in any model developed of the process. Four models of clinical judgment were found in the literature (Gordon, Murphy, Candee, & Hiltunen, 1994; Kataoka-Yahiro & Saylor, 1994; Pesut & Herman, 1998; Tanner, 2000). The Tanner (2000) research-based model of clinical judgment was chosen as the foundation for this study. Each of these models will be explored to identify key features of clinical judgment and critiqued in analysis of their strengths and limitations.

The Integrated Model of Clinical Judgment

Gordon, Murphy, Candee and Hiltunen presented their integrated model of clinical judgment in 1994. This model is diagramed around a central core of the nursing process. The two domains of clinical judgment, identified as diagnostic-therapeutic and ethical, flow down either side mirroring the generic core nursing process. The absence of a clearly stated definition of clinical judgment makes analysis of this model difficult. However, the authors bring forward several points of interest. First, in discussion of information collection and interpretation, they note the focus for information analysis will reflect the interests of the nurse. Further, they state nurses' philosophical basis for practice, beliefs about the conceptual focus for nursing and its social mandate and value system will impact the scope of information collected and its interpretation. They also note that nurses use different reasoning patterns based on their knowledge and skill and sensitivity to cues, and note these reasoning patterns can vary between hypothetical deduction to nonanalytic pattern recognition. These issues are important because they represent areas of consistency between Gorden and colleagues (1994) and Tanner's research-based model of clinical judgment.

However, the integrated model of clinical judgment produced by Gordon and colleagues (1994) was rejected for several reasons. First, they critique the separation of ethical decision making from the process of diagnostic and therapeutic decision making. They state diagnostic and treatment decisions are rarely made without reference to values and that most ethical concerns emerge from the context of nursing. They state both domains use the same reasoning processes and only differ in content and criticize the absence of an ethical component in the nursing process. Yet, while they state they are presenting an integrative model of clinical judgment to resolve this issue, on close examination, the model remains a depiction of parallel processes.

A second criticism of Gordon et al.'s integrated model of clinical judgment relates to their use of the nursing process as the central core of their model. The authors state clearly that although the model represents clinical judgment as a linear process, it is not. They state clinical judgment is complex and iterative and difficult to capture in a flow diagram. The discontinuity between the description of their model and the figure of the model suggests the concept domain of clinical judgment is not fully articulated. The absence of a definition of clinical judgment, along with their varying uses of the terms clinical judgment, clinical reasoning and clinical decision making, further support this position.

A final criticism of this model is that it presents a static view of clinical judgment. As it is built around a core of the nursing process, it begins with the recognition of a problem and ends with an evaluation of the outcome attained. While a strength of the model is their recognition of the nurse as a powerful moral agent, there is no mechanism in this model that demonstrates how the application of the process impacts the ongoing development of the nurse.

The Critical Thinking Model for Nursing Judgment

Kataoka-Yahiro and Saylor (1994) developed a critical thinking model for nursing judgment. These authors identify clinical judgment as an outcome of critical thinking.

Their model describes the components of critical thinking to include content knowledge, experience, competencies, attitudes and standards. They link critical thinking to nursing judgment in a developmental manner. Three levels of intellectual and ethical development are identified – basic, complex and commitment. As the nurses' critical thinking abilities become more refined, they are able to see more alternatives in a situation. The goal in the model is attainment of the topmost, or commitment level, of critical thinking in nursing.

Kataoka-Yahiro and Saylor differentiate between nursing judgment and clinical judgment. They define nursing judgment as the outcomes of decisions made using critical thinking. They differentiate the types of decisions made by the analysis of the role of the nurse in the situation. Direct nursing judgments are those made about direct patient care provided by a nurse. Semi-direct nursing judgments are those made by directors of nursing about distribution of resources. Indirect nursing judgments are those made by nurse educators when making curricular decisions. Clinical judgment, in the authors' view, was limited to an exclusively direct care situation. This model was not used in this proposal because the focus was on delineating critical thinking processes as opposed to the processes of clinical judgment, which was essential to this research.

The Outcomes Present State Model of Clinical Reasoning

Pesut and Herman (1998) developed the outcome present state test (OPT) model of clinical reasoning. They trace the development and modifications of the nursing process from its inception in the 1950s. They identify their model as a third generation of the nursing process developed in response to the shift in nursing from problem identification and diagnosis to outcomes assessment. The OPT Model represents a significant shift from its historical antecedents. The model initiates with the client-in-context story, which provides the facts and cues that are linked in logical ways to develop a framework for reflection and decision making. Within the frame, the patient's present state is juxtaposed with outcome states to establish a match or mismatch test condition. Clinical judgments are the conclusions identified from the tests comparing present state to outcomes state. Clinical decisions are the choice of nursing actions taken to help the transition from the present state to the outcome state. Clinical judgments are iterative depending on the meaning of test results.

The OPT is a strong model of clinical reasoning but it truly is a third generation nursing process model. It is designed to explain how metacognition impacts the nurses' ability to diagnose, intervene and evaluate outcomes. Kautz, Kuiper, Pesut, Knight-Brown and Daneker (2005) conducted a study to examine the effect of learning strategies based on the OPT model, with the addition of a reflective journal activity based on Kuiper's (2002) self-regulated learning theory, on the development of clinical reasoning skill. While the study was limited by the absence of a control group, they did find that students did improve in their ability to frame a clinical situation and make decisions regarding appropriate interventions.

The OPT Model as a framework for this study has important limitations. With the nursing process core, the model is not designed to explain how nurses make choices in the moment and balance competing priorities. Thus, it does not support a performance assessment application. Also, the nursing process core to the model does not support the comparison of the differential reasoning patterns used by expert nurses to those of novice nurses. Rather, the model seems most useful in making explicit the clinical reasoning processes of novice nurses.

The Research-Based Model of Clinical Judgment

The research-based model of clinical judgment as presented in Figure 1 (Tanner, 2006) was selected to provide the theoretical foundation for this study for several reasons. First, this model emerges from the strongest synthesis of literature concerning clinical judgment. In the most recent articulation of the model, Tanner (2006) identified 191 research studies were synthesized in order to derive this model. No other model of clinical judgment brings forward this level of evidentiary support. A second advantage of this model, that is not present in the prior models, is the explicit focus on clinical judgment. Tanner rejects the nursing process as an explanation of how nurses think. Instead, the model presents a view of nursing that reflects the contextual process of how a nurse responds to the conflicting priorities involved in patient situations. A third advantage of this model is the view that clinical judgment is developmental. The reflective capability of each nurse facilitates the carryover of knowledge from the present to future situations. These advantages highlight some of the important differences between the research-based model of clinical judgment and the earlier models. Given the importance of this model as the theoretical foundation for this study, it will be examined in detail.

The research synthesis conducted by Tanner (2006) identified three factors that influence subsequent clinical judgments (1) the nurses' background related to the specific situation, (2) the nurses' relationship with the patient and (3) the context of the interaction. The background for nurses' practice consists of theoretical and practical knowledge as identified by Benner (1984). Theoretical knowledge or “knowing that” consists of formal abstract rules and principles that emerge from scientific research and acquired from textbooks and classroom lecture. Practical knowledge or “knowing how” is gained from experience and is highly context dependent. The nurses’ background also contains their personal values, attitudes and beliefs, which can form the basis for biases, blind spots, or preconceptions (McCarthy, 2003). The final component of the

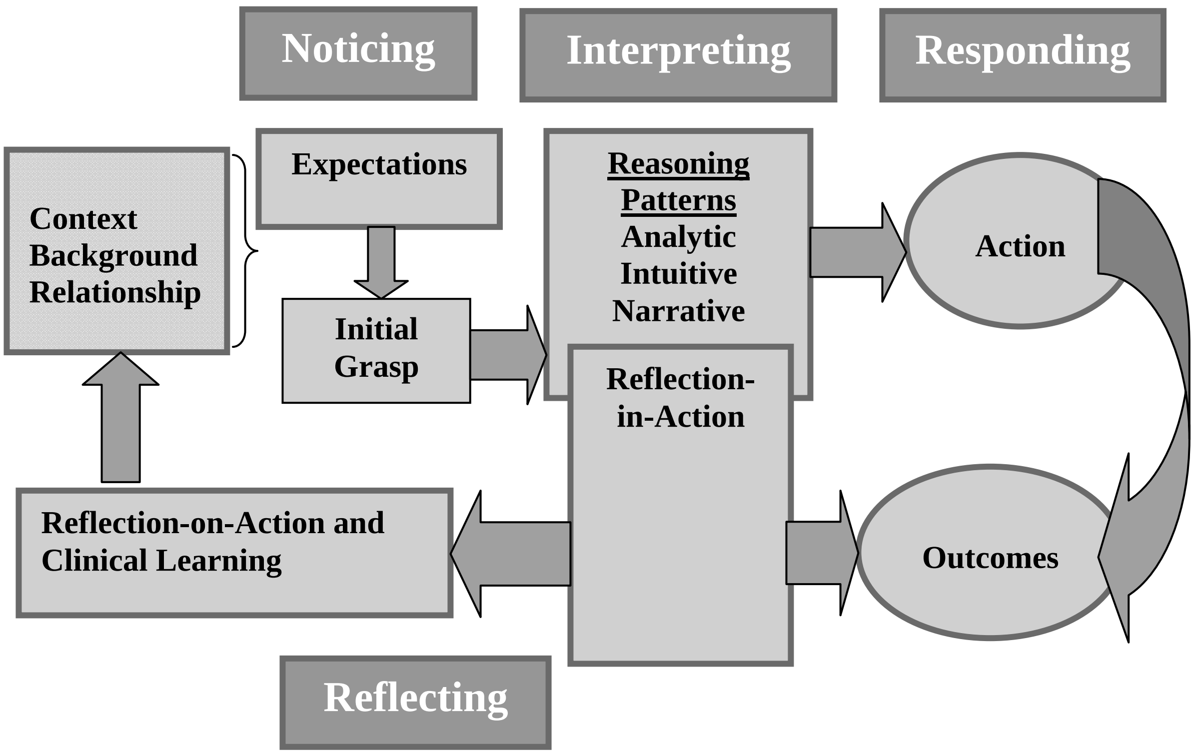


Figure 1. The Research-Based Model of Clinical Judgment, 2006

background for practice is the nurses’ conception of excellent practice. Their disposition toward what is good and right is socially constructed and embedded both within the discipline and within the norms and mores of the particular group with which they practice (Benner, Tanner & Chesla, 1996).

The second influence on clinical judgment is the nurses' relationship with the patient (Tanner, 2006). This relationship is commonly referred to by nurses as 'knowing the patient' and is described as a tacit, taken-for-granted understanding of their patients (Jenks, 1993; Jenny & Logan, 1992, Tanner, Benner, Chesla & Gordon, 1993). Tanner et al. (1993) describes two different types of knowing. One type of knowing reflects an understanding of the patient's typical pattern of responses, which allows for the development of a sense of salience. Salience is the ability to focus on the most relevant cues in the current situation. The second type of knowing reflects an understanding of the patient as a person. Knowing the patient as a person allows nurses to correctly judge their level of involvement, which influences not only engagement in problem solving but also the outcome and the nurses' sense of satisfaction (Benner et al., 1996; Morse, 1992, Tanner et al, 1993).

The third influence on clinical judgment is the context of the situation and the culture of the nursing unit (Tanner, 2006). The work group itself, the habits and culture generated, determine the knowledge valued, the skills taught and the situations that require judgment (Benner, Tanner, Chesla, 1996; Ebright, Patterson, Chalko & Render, 2003; Ebright, Urden, Patterson & Chalko, 2004). Further, power and status inequities between nurses and other providers have been found to impact nursing clinical judgment by influencing both the nurses desire to seek understanding of problems and their ability to intervene therapeutically (Benner, Tanner & Chesla, 1996; Bucknall & Thomas, 1997).

To summarize, three factors, the nurses' background, their relationship with the patient and the context of the situation all influence the process of clinical judgment. The importance of situational factors to the subsequent clinical judgment is a position held in common across three of the four models of clinical judgment, the integrated model of clinical judgment by Gordon et al (1994), the OPT Model by Pesut and Herman (1998) and the research-based model of clinical judgment by Tanner. The commonality of this viewpoint emphasizes its importance to the construct of clinical judgment.

The situational backdrop sets up the nurses' expectations for each patient encounter. The nurses' expectations provide the framework for their initial grasp of the situation. Noticing is the term Tanner (2006) uses to define the nurse's initial perceptual grasp of the situation. Noticing is not a culmination of conclusions from assessment data. Noticing is the result of the nurses' ability to compare their expectations with their actual findings. This ability to notice will vary as a result of the previously identified factors that interact to influence the situational backdrop. Noticing inherently initiates the clinical judgment process; nurses cannot interpret and respond to data that are not noticed.

The next component of clinical judgment is interpreting. Interpreting describes the process nurses use to make sense of the data that were noticed and establish priorities. From the evidence in the literature on decision-making, it is clear that nurses use many different types of reasoning patterns depending on the nature of the task and the context (Brannon and Carson, 2003; Lauri & Salantera, 1998; Simmons et al. 2003). Commonly used forms of clinical reasoning include analytic processing, intuition, and narrative thinking. The most common type of analytic reasoning is hypothetical-deductive, a process undertaken if there is a mismatch between what was expected and what was actually found, or if essential knowledge is lacking or if there are multiple options available. This analytic process is characterized by the generation of alternative choices and the conscious weighing of the data against the possible outcomes (Tanner, 2006). Intuition as a reasoning pattern is characterized by an immediate grasp of the situation and typically improves with experience in similar situations. Intuition has become recognized as a form of pattern recognition (Benner, Tanner & Chesla, 1996). In narrative reasoning, understanding of the particular, is acquired through story telling and interpretation, and used to acquire understanding of the general (Bruner, 1986). Narrative thinking helps the nurse acquire an understanding of the meaning behind the experience, which facilitates their understanding of the patient as a person. Understanding the specific case also helps convert experience into practical knowledge. The key point with interpretation as part of the clinical judgment process is the realization that all nurses use all types of reasoning processes at different time and in different situations (Tanner, 2006).

Responding represents the clinical judgment component of taking action and evaluation of outcomes. Effective responding requires reflection-in-action and this aspect differentiates clinical judgment from the nursing process. The nursing process is a method by which generic nursing interventions are attached to identified nursing diagnoses and applied in specified situations. Clinical judgments are contextually bound responses to a particular patient's healthcare alteration. The actions taken as a result of a clinical judgment, or the choice to take no action, are in response to the patient's situation. The actions taken as a result of the nursing process are in response to the generic diagnosis identified. Reflection-in-action allows the nurse to bring forward what is known about the patient and about the situation in order to best determine a course of action.

Once an action is taken in response to a health problem or concern, the outcomes that result can be evaluated. Reflection on outcomes can be cursory or thorough. Mezirow (1991) identified three categories of reflection in adult learners; the non-reflector, reflector and critical reflectors. Non-reflectors view the situation impersonally with little awareness of context. Reflectors and critical reflectors are able to examine the situation at increasingly deep levels. If reflection occurs, it becomes a link between the experience and the development of expertise (Bransford, 2000; Kuiper & Pesut, 2004). By linking reflection back to the nurse's background that is carried forward to the next situational encounter, the research-based model of clinical judgment becomes dynamic. The dynamic nature of the model supports its applicability as a framework to describe clinical judgment across the spectrum from novice to expert nurse.

In summary, four models of clinical judgment were identified in the literature. From these four models limited research was located using these models as a foundation for further investigation. No studies were found using the integrated model of clinical judgment from Gordon and colleagues (1994). One study was found using the Kataoka-Yahiro critical thinking model of nursing judgment (Ellermann, Kataoka-Yahiro & Wong, 2006). This study examined the use of various logic models within a curriculum and their impact on concept maps written by students. Pesut's most recent work using the OPT model (Kautz, et al., 2005) focused on assessment of students' written work and clinical faculty applied a dichotomous rating of 'evident' or 'not evident' to the components of the OPT. Thus the ongoing research from these models was limited to cognitive appraisals of student ability in clinical judgment. From the research-based model of clinical judgment, one study was found (Lasater, 2005) using this model as the theoretical foundation for the development of a clinical judgment performance evaluation. The focus of this present study is a further examination of the construct validity of the instrument developed by Lasater. However, nurses have historically evaluated clinical judgment using a variety of methods and theoretical foundations, and these options will be critiqued next.

Methods of Measuring Clinical Judgment Ability

Critical Thinking Assessments

The critical thinking literature in nursing has been plagued with multiple definitions for this construct (Gordon, 2000; Jones & Brown, 1991). This lack of definitional clarity impacted research evaluation, as each definition seemed to generate a separate measurement instrument (Rane-Szostak & Robertson, 1996). Three measures in particular have been widely used in nursing research. In early research the most widely used measure was the Watson-Glaser Critical Thinking Appraisal (WGCTA). In the more current research, the instruments used were the California Critical Thinking Skills Test (CCTST) and the California Critical Thinking Dispositions Inventory (CCTDI), both developed from the work of Peter and Noreen Facione (1990; 1995; 1996). The literature pertaining to critical thinking will be limited to a review of the findings related to these three instruments.

Critical thinking in nursing, as measured by the WGCTA, has been the subject of multiple integrated reviews (Adams, 1999; Adams, Whitlow, Stover & Johnson, 1996; Hicks, 2001; Kintgen-Andrews, 1991). Hicks (2001) summarized the decade of research concerning the WGCTA with three conclusions. First, students did not consistently show improvement in WGCTA scores as a result of nursing education. Second, the WGCTA did not consistently reveal differences in critical thinking ability across different types of nursing programs. Third, the WGCTA was inconsistent with other measures of clinical judgment and decision-making. Possible etiologies for these inconsistencies were identified as either design flaws in the instrument or in the studies using the instruments (Hicks, 2001) or that nurses use abilities other than those measured by general critical thinking instruments (Adams, 1999)

From the late 1990s forward the primary instruments used to assess changes in student critical thinking abilities have been the California Critical Thinking Skills Test and the California Critical Thinking Dispositions Inventory. These two separate instruments were developed by the Facione’s and reflect their definition of critical thinking as six separate cognitive skills and six separate dispositions (1996). Eight studies were found using these two instruments with undergraduate nursing students and were clustered for analysis by research design. Two studies examined critical thinking across program type. Spelic, Parson, Hercinger, Andrews, Parks and Norris (2001) conducted a longitudinal study of traditional, accelerated and RN to BS students and found a significant improvement in CCTST scores from program entry to exit. Shin, Jung, Shin & Kim (2006) compared associate degree with traditional baccalaureate and RN to BS students and found baccalaureate students scored significantly higher than either the associate degree or RN to BS students on both the CCTST and the CCTDI. Three studies examined critical thinking using cross-sectional samples of baccalaureate students. Colucciello (1997) using the CCTST found first semester juniors scored highest followed by first semester and then second semester seniors reflecting a pattern of critical thinking skills that diminishes with nursing education. McCarthy, Schuster, Zeha and McDougal (1999) found that scores on both the CCTST and the CCTDI improved from sophomore to senior year. Profetto-McGrath (2003) found that while CCTST scores did increase and the CCTDI did show students had a positive inclination toward critical thinking there was no significant difference found from freshman to senior groups. Two longitudinal studies were found. Thompson and Rebeschi (1999) followed one cohort of students for a two-year period and found significant increases in CCTST and CCTDI scores. Beckie, Lowry and Barnett (2001) followed three cohorts of students, their last class of an old curriculum and two classes from a newly designed curriculum. They found the first year of the new curriculum students achieved significantly higher CCTST scores but the second class failed to demonstrate similar scores.

These eight studies summarize a pattern of findings on the CCTST and CCTDI that is similar to that found when critical thinking was studied using the WGCTA. Nursing education does not consistently show improvement in student critical thinking ability when assessed using a general measure of critical thinking ability. The absence of a nursing specific definition of critical thinking was suggested as a source of these inconsistent findings (Brunt, 2005; Rane-Szostak & Robertson, 1996). The thought was that definitional clarity would promote alignment between research instrumentation and measurement. However, Scheffer and Rubenfeld developed a nursing specific definition of critical thinking in 2000 but there has been no subsequent development of a nursing specific measure of critical thinking. The inconsistencies of the research findings on the general measures of critical thinking along with the continued program outcome goals regarding critical thinking held by the AACN (1998) and the NLNAC (2005) keep the search for an effective measure of critical thinking ongoing.

The focus of current literature has been the relationship between measures of general critical thinking and nursing clinical judgment. There have been several recent integrative reviews of this literature (Brunt, 2005; Hicks, 2001; Staib, 2003). The summary findings from these reviews indicate there is no clear or consistent relationship between general critical thinking ability and clinical judgment. These findings in conjunction with the difficulties in measuring critical thinking have nursing leaders questioning our priorities toward assessment of critical thinking (Riddell, 2007; Tanner, 1998). Scheffer and Rubenfeld (2007) recommend that critical thinking be conceptualized as a tool for developing the expert nursing judgment needed to improve patient outcomes; that critical thinking needs to be placed in a nursing context to be useful. Walsh and Seldomridge (2006) have similar recommendations to contextualize critical thinking. However, they go further to recommend that critical thinking be deconstructed to extract those aspects most relevant for nursing and those be defined and operationalized for different levels of students. In addition Walsh and Seldomridge (2006) suggest that the clinical setting is the priority location for both the teaching and the assessment of critical thinking.

This summary of literature pertaining to critical thinking has several implications for a study of the construct validity of a clinical judgment rubric. First, one source of evidence for construct validity is that based on the relationship between the construct of interest and other variables. This summary of literature suggests that general measures of critical thinking ability such as the WGCTA or the CCTST or the CCTDI would not be effective sources of comparative variables in a validity study of clinical judgment.

Second, the inconsistencies in the studies examining the impact of nursing education on critical thinking ability and the lack of a relationship between critical thinking and clinical judgment support the position that critical thinking is not the central construct determining nursing ability. Third, the current recommendations to improve critical thinking education and research emphasize placing it in a clinical context. Thinking like a nurse, making choices in a clinical setting, defines clinical judgment and supports this construct rather than critical thinking, as most central to determining nursing ability.

The Jenkins's Clinical Decision Making in Nursing Scale

In 1985 Jenkins's developed the Clinical Decision Making in Nursing Scale. The instrument was designed to examine nursing students' self-perceptions of decision making with the goal of using the data to examine curricular implications. The instrument was based on the work of Janis and Mann (1977, as cited in Jenkins, 1985) and their seven criteria needed to elicit a state of "vigilant information processing" were condensed into four categories of decision making: (1) search for alternatives or options (2) canvassing of objectives and values (3) evaluation and reevaluation of consequences and (4) search for information and unbiased assimilation of new information. Decision making in this study was defined as "a conscious, cognitive impression of how one goes about making decisions" (Jenkins, 1985, p 222).

The instrument that emerged from rigorous content and reliability evaluation consisted of 40-items rated by nursing students from 5 (always) to 1 (never) that reflected their perceptions of their behavior while caring for patients. Low scores represented a negative perception of decision making while high scores represented a positive perceptions. Jenkins' initial study used a cross-sectional sample of sophomores, juniors and senior nursing students. Jenkins hypothesized that decision making scores would increase as students progressed through the nursing program as they acquired more experience and skill. However, the study found no significant differences among the three student groups except for the first category, search for alternatives or options. In this category, seniors scored significantly higher than juniors but sophomores did not differ significantly from either the senior or the junior groups.

As a measure of clinical judgment ability, both its design and its theoretical foundations limit the Jenkins Clinical Decision Making in Nursing Scale. First, the instrument was designed as a measure of the self-perception of decision making; Jenkins clearly excluded observation of student decision making in the classroom or the clinical setting from the study. Thus, the scale as originally designed is not applicable for use as a clinical performance assessment. Further, the theoretical foundation of the study itself is limited by reliance on information processing theory. Jenkins used a rational, normative viewpoint to define decision making and to frame the instrument, and later identified this as a limitation (2001) agreeing “some basis exists for the presumption that total rationality is not possible in the real world” (p36). Nurses’ problem solve under the influences of time and the situation, both of which limit the information available and their ability to process it.

Patient Management Problems

To move clinical judgment performance appraisal beyond the limitations of a self-perception measure, researchers turned to simulation. Simulation is an educational method where aspects of a real-world domain are artificially abstracted and replicated in order to achieve a defined learning goal (Gaba, 2004; Hertel & Millis, 2002). The term fidelity is frequently used in conjunction with simulation and the level of fidelity refers to how closely the simulation replicates reality. Fidelity is determined by both the number of elements of reality that are replicated as well as the amount of error allowed between each element in the simulation and the real world (Gaba, 2004). Simulation can be enacted many ways, from paper-and-pencil to computer-assisted scenarios through live actor role-playing to a patient manikin.

Patient management problems (PMPs) were an early simulation prototype (Schleutermann, Holzemer, & Farrand, 1983). Paper-and-pencil branched simulations were designed with an initial brief description of the scenario and learners were required to make judgments about what data to collect, how to interpret findings and subsequently manage the care of the patient. Selections were made from a large number of options and the pathway through the problem varied with choices made. Scoring compared the choices of the participant with that of an expert panel; each item was considered essential, contributory but not essential, or inappropriate. The resulting proficiency score compared the pathway of the participant to the optimum pathway and was reported as a percentage.

Psychometric evaluation of PMPs conducted by Holzemer and associates (Farrand, Holzemer & Schleutermann, 1982; Holzemer, Resnik & Slichter, 1986; Holzemer & McLaughlin, 1988) used a national sample of nurses. The 1982 study supported the construct validity of the PMP by demonstrating the measure could discriminate between the performance of nurses (associate degree, diploma and baccalaureate level) and the performance of nurse practitioners on three separate scenarios. However, the 1986 study that examined the evidence based on relationships to other variables, or as stated in the study the criterion-related validity, was less successful. The performance on PMP simulation was compared to performance in actual practice as represented by chart audit and direct observation. The study found no correlation between the performance on the PMP and either the chart audit or the direct observations. As a result, neither the concurrent nor predictive aspects of construct validity of the PMP as a performance measure were supported. Holzemer and colleagues suggested the lack of criterion-related validity for PMPs could be partially explained by the limited fidelity of the PMP and the then current inadequate understanding of clinical problem-solving. Clinical judgment is domain-specific and contextual. When confronted with contextual variation, the PMP could not discriminate among performance levels.

The Lasater Clinical Judgment Rubric

Subsequent changes in our understanding of the clinical judgment process and in the level of technology available in simulation addressed the methodological deficits present in the PMPs. The research-based model of clinical judgment represents a significant forward step in understanding the process of clinical problem solving. The new generation of high-fidelity patient manikins represents a significant step forward in addressing the difficulties of contextual variation. The new manikins are capable of interactive speech as well as portraying physiological assessment data, which can be altered in response to the nurse's actions. The use of high-fidelity simulation allows educators to control the practice environment. The setting in high-fidelity simulation is realistic but repeatable allowing the presentation of basically equivalent situations to multiple participants. These developments were used to advantage by Dr. Lasater in the building of a clinical judgment evaluation rubric.

The Lasater Clinical Judgment Rubric (Appendix A) is an instrument that describes the four dimensions of clinical judgment, noticing, interpreting, responding and reflecting in specific detail. The four dimensions of clinical judgment become four subscales with each containing specific indicators. The dimension of noticing is defined by the nurses' ability to conduct focused observation, recognize deviations from expected patterns and seek information. The dimension of interpreting is defined by the ability to set priorities and make sense of data. The dimension of responding is defined by the nurses' calm, confident manner, clear communication, well-planned interventions and flexibility, and skillfulness. The dimension of reflecting is defined by ability in evaluation and self-analysis along with commitment to improvement. Four levels of ability are identified as beginning, developing, accomplished and exemplary. Thus the rubric presents as an eleven by four table. The goal for each of the descriptors in the rubric was to identify behaviors and verbalizations indicative of a student's understanding and ability and/or non-understanding and inability in the clinical judgment process (Lasater, 2005).

Development of the clinical judgment rubric was a lengthy process. Beginning with a strong foundation in the literature behind the clinical judgment process, Dr. Lasater used observations of students encountering clinical problems in a high-fidelity simulation theater. As students demonstrated differing levels of clinical judgment ability in response to the same clinical problem, Lasater was able to use these observations to develop behavioral descriptors. In addition, in the process of rubric development Lasater consulted frequently with Dr. Tanner, developer of the clinical judgment model, as well as two others, an educational specialist in rubric design, and the simulation facilitator.

The goal of these meetings was to refine the language used to describe student behaviors and clarify the differences between the dimensions and levels of clinical judgment.

The actions taken by Dr. Lasater during rubric development support the construct validity based on content. The situational consistency available in high-fidelity simulation provided a stable base to reveal the differing levels of student ability in response to the same test construct. The use of a highly qualified and diverse group of consultants to evaluate the representativeness of the rubric descriptors developed by Dr. Lasater from the observational process supports the evidence for construct validity based on content.

The Lasater rubric was designed for quantitative measurement of clinical judgment performance. Behaviors in each dimension are scored according to level, with novice behaviors scored as a one and exemplary behaviors scored as a four. The preliminary work (Lasater, 2005) assessing the ability of the rubric to define a student's stage of clinical judgment development was conducted in the high-fidelity simulation setting. The rubric was applied to a group of 26 students enrolled in their first adult acute care medical-surgical nursing course. Student observations were conducted during both the active component of the scenario and during debriefing. The mean rubric score was 22.98 (SD = 6.07) placing the group average in the developing stage, which was the finding expected given the students' educational level. The main limitation of this finding was the researcher, prior to scoring, knew the current education level of the students. While the use of behavioral descriptors and direct observation of student performance minimizes this limitation, a further validity study is needed.

Summary of Literature Regarding Clinical Judgment

The purpose of this study was to further examine the construct validity of the clinical judgment evaluation instrument developed by Lasater (2005). This study used known-groups methodology to examine the evidence based on relationships to other variables. Using the logic of known-groups, if the Lasater Clinical Judgment Rubric was sensitive to differences in clinical judgment known to exist between the groups, this finding supports the construct validity of the measure. Central to the use of known-groups methodology was establishing the two groups as different on the construct of interest. In this study the hypothesis tested was that graduating seniors would demonstrate a higher level of clinical judgment than end-of-year juniors. Therefore, it was important to establish that these two groups do possess differing levels of clinical judgment. The next section of the literature review summarizes the research addressing the impact of education on each of the components of the clinical judgment process.

The Literature Pertaining to Noticing

In 1968 Verhonic, Nichols, Glor and McCarthy designed a study to identify how a nuance observed by a nurse would affect the action taken. The study was built on the premise that the ability to observe, evaluate, interpret and take action was the primary professional responsibility of the nurse. Five film clips were developed, each lasting one to two minutes. Each film clip was preceded by a short introduction. After watching the film, study participants answered three questions: 1) what did you observe? 2) what action would you take based on what you saw? 3) what led you to take this action? The initial study analyzed responses from 1,576 participants. Observations were coded as relevant, irrelevant or inappropriate. With regard to relevant observations, their findings demonstrated that as educational level increased, the number of relevant observations made also increased. This study was replicated twice (Davis, 1972; 1974). The first replication study compared two groups, clinical nurse specialists (CNSs) and baccalaureate nurses (BSNs). In the second study three groups were compared: CNSs, BSNs and diploma graduates. Both of the Davis studies had similar findings with regard to the positive impact of education on the ability of the nurse to notice relevant cues.

The findings of Verhonic, Nichols, Glor and McCarthy (1968) and Davis (1972; 1974) regarding the impact of years of nursing experience on the ability to take action are more complex. In all three studies, actions were coded as therapeutic (dependent upon an order from a physician), supportive (independent nursing action) or inappropriate. The studies found nurses without a degree recommended the highest number of therapeutic actions and least number of supportive actions. Nurses with doctoral degrees listed the least number of therapeutic actions and were second to the master’s degree group in frequency of supportive actions. All three studies found that as years of experience increased, the number of observations, actions and reasons decreased. Verhonic and colleagues recommended that the relationship be studied further. Davis went on to state years of experience diminished the quality and quantity of nursing care provided. Given the design of these studies, this conclusion is not supported. The research design only addresses volume of observations and dependent or independent nursing actions in response; the quality component of relevance or prioritization was not included. An alternative explanation for the decrease in the number of observations, actions and reasons cited by experienced nurses in these studies could be their perceptual grasp of the relevant cues is better than those with less experience as is their ability to identify the priority action needed to effectively respond. Thus, the data from these studies does not disconfirm the positive relationship between experience and clinical judgment.

An additional study used the Verhonic and colleagues film clips to examine the impact of education. The Frederickson and Mayer (1977) study is commonly reported (Bowles, 2000; Kintgen-Andrews, 1991; Pardue, 1987) as supporting the position that education does not improve problem-solving. Closer examination of the study design does not support this conclusion. Frederickson and Mayer used three of the five film clips developed by Verhonic and colleagues to compare the problem-solving abilities of baccalaureate nurses with associate degree nurses. The sequences elected were a young woman scheduled for a radical mastectomy, a man experiencing symptoms of an acute myocardial infarction, and an elderly man who falls out of bed. After viewing the film clip each student was asked to think aloud and their responses were evaluated against a faculty defined problem solving criteria. All students also completed a standardized test of general problem-solving ability. A weakness of the study is the failure to report statistical findings; however, the researchers report no significant differences in problem solving between BSNs and ADNs nursing students. They did find BSNs scored significantly higher on an unidentified standardized test in critical thinking.

The design flaws in the Frederickson and Mayer study emerge from the misconception of problem-solving as a static rather than a dynamic process. The authors wonder why students infrequently address the step of evaluation in their problem solving and cannot explain why BSN students scored higher in critical thinking when they were not also better at problem solving. One solution is that the students viewed the scenarios as a static assessment of their theoretical knowledge and not as representative of a dynamic problem and the knowledge assessed by the problems presented in the film clips was foundational for both programs thus a ceiling effect precluded findings differences between the two groups.

Itano (1989) analyzed the clinical judgment process of highly skilled nurses in comparison to the process exhibited by student nurses. The method involved audio-taping and directly observing study participants as they initially assessed a patient. The interview was then replayed and the nurse was asked to review her “thoughts” with the researcher, the description of which was also audio taped. The two audiotape transcripts were coded for four types of clinical cues and evaluated for level of clinical judgment exhibited on a researcher designed clinical judgment rating scale. Itano found that highly skilled nurses collected more cues about their patients than student nurses, a finding consistent with Verhonic and Davis. Further, the study found a significant difference between the highly skilled nurses and the student nurses on the clinical judgment process rating scale. This study supports the influence of education and experience on the development of clinical judgment.

The clinical judgment process rating scale designed by Itano shows some similarities with the Lasater rubric but is limited by its focus on only the process of diagnosing. The instrument examines sources of data and how the nurse organizes information to develop and evaluate alternative hypotheses. The actions of responding to the information and reflecting on outcomes are missing. The scale used by the instrument was a 5-inch line with two anchors, novice ratings were at the left end and highly skilled ratings were at the right end. This clinical judgment rating scale represents an early attempt to differentiate level of performance. The ability of the instrument to discriminate between student nurses and highly skilled nurses in the areas of noticing and cue interpretation provides early support for the model of clinical judgment.

Reischman and Yarandi (2002) also studied diagnostic accuracy and cue utilization in expert and novice nurses. Four written simulations were used in data analysis. After reading the simulation, the nurse was asked to verbally recall the simulation, identify the primary problem and provide a therapeutic rationale. Diagnostic accuracy and cue utilization scores were determined by a comparison between the participants' reports and those provided from consensus of an expert panel. This study found that expert nurses were significantly more accurate in diagnoses than novice nurses. While the novice nurses were able to recall significantly more total cues than experts, the expert nurses were able to recall a significantly larger number of highly relevant cues than novices. These findings support the position that experience does influence noticing ability. The acquisition of domain-specific knowledge that comes with experience facilitates the ability of experts to focus on highly relevant cues and demonstrate a sense of salience, when compared to novice nurses.

This set of five studies consistently reveals a pattern that education and experience do influence the ability of nurses to notice cues. The early studies of Verhonic et al (1968) and Davis (1972, 1974) when compared to the work of Itano (1989) and Reischman and Yarandi (2002) differ in specificity of findings. The early work stating that experts' notice more cues than novices and the later work stating the difference lies in level relevance of the cues noticed. This difference can be attributed to an improvement in the understanding of the cognitive processes associated with noticing as well as to refinements in study design. The consistent finding among all five studies is the relationship between domain-specific knowledge and ability to notice. As domain-specific knowledge increases, either through education or experience, noticing becomes more refined.

The Literature Pertaining to Interpreting & Responding

The literature addressing the development of the clinical judgment abilities in the dimensions of interpreting and responding must be extracted from the literature describing decision-making. Decision-making is a cognitive process resulting in the selection of a course of action from alternatives. There are at least three theories of clinical decision-making, each using different assumptions and research methods. The research pertaining to the clinical judgment dimension of interpreting is embedded in the literature of diagnostic reasoning. The research pertaining to the dimension of responding is found throughout the decision-making literature. These differences make it difficult to compare the findings of different studies. However, if the summary is narrowed to focus on the goal of showing that the skills of novice nurses at interpreting data and responding are different from that of expert nurses, and focused on showing how education and experience impact the development of these skills, some conclusions can be identified.

Any discussion regarding how the decision making capacity of a nurse changes over the course of a career must begin by acknowledging the contributions of Benner. In 2004 she published a summary of almost 30 years of research on her model of the stages of skill acquisition in nursing practice and reported the model to be “predictive and descriptive of distinct stages” (p.198). One of the hallmarks of an expert, according to Benner (2004) is a rapid and integrated response. In the expert, the language of recognition and assessment are limited because they are so linked with actions they have become self-evident. The expert has a rich sense of expectations but also a rich sense of engagement with the situation; this position leaves the expert open to notice the unexpected and accept evidence that disconfirms assumptions. An early study by Corcoran (1986) showed that expert nurses when dealing with complex patients were able to generate more alternative actions, provide greater depth of information given to patients and more critically evaluate their care than novices.

At the other end of the spectrum, Benner (2004) located the novice within the first year of education. The novice is described as one who enters a clinical situation without any background to form a basis for understanding. Rules and guidelines are necessary for the novice to begin a safe, situated learning experience. Novices need to be coached to compare and contrast textbook examples with actual clinical cases; exceptions and contradictions to the norm must be identified for the novice. Much of what the novice learns is practical knowledge where understanding is attained only through experience with patients.

Research has confirmed this view of the novice. Thiele, Holloway, Murphy, Pendarvis and Stucky (1991) studied 83 baccalaureate nursing students at the end of their first clinical course. Actual decision making was determined by student performance on a paper-and-pencil simulation, which used multiple choice and short answer questions to elicit students' views of relevant cues, priority nursing diagnosis and interventions. Perceived decision making was measured using the Clinical Decision Making in Nursing Scale by Jenkins (1985). The study found novice students are incapable of differentiating relevant from irrelevant cues; all cues are considered of equal importance. Tschikota

(1993) studied novice nurses with similar results. She found novices did generate hypotheses but based their decisions on fact. Students were capable of matching factual cues to their theoretical knowledge base. Tschikota (1993) concluded the “better knowledge base a student has, the more likely that student will be to choose data that is pertinent to decision making” (p. 396). These two studies suggest that a difference between the novice and the expert is their ability to make sense of the data as it relates to a specific patient and to set appropriate priorities.

The difficulty novice nurses have in determining cue relevance increases with the complexity of the clinical situation. Distinct differences between novices and experienced nurses are apparent when the situation becomes more ambiguous (Brannon & Carson, 2003; Tabak, Bar-Tal, & Cohen-Mansfield, 1996). In situations when the information is structured and consistent, novices and experienced nurses are both certain of their decisions. In situations where the information is inconsistent and unstructured, experienced nurses are appropriately less certain of their decisions. Experienced nurses are more capable of recognizing ambiguity and changing their diagnostic approach. Novice nurses are less likely to recognize the implications of the information that disconfirms their expectations. Brannon & Carson (2003) suggest that novices may apply a heuristic process and decide the fit is “close enough” and disregard the inconsistencies.

Making clinical decisions in conditions of ambiguity, inconsistency and complexity requires domain-specific knowledge. The ability of a nursing program to provide that domain-specific knowledge was the focus of a study by Botti and Reeves (2003). They studied nursing students with high and low academic ability at two levels of domain-specific knowledge. The first group had completed two full years of accompanied clinical practice and the second group had completed three full years of accompanied clinical practice. Decision making ability was determined by performance on paper-and-pencil simulations of varying complexity from easy to difficult to impossible. The purpose of the study was to determine if there was a difference in the students' ability to generate alternative hypotheses, to identify disconfirming information, to recognize the need for more information and diagnostic accuracy.

Botti and Reeves (2003) found with easy clinical problems, the only significant variable in diagnostic accuracy was academic ability. Students with high academic ability, irrespective of year, made more accurate diagnoses when the complexity of the task did not require interpreting information that could lead to alternative diagnoses. However, as the complexity of the clinical problem increased the variable that differentiated between student performance was years of study. Third year students were significantly better than second year students in their ability to select disconfirming cues that were relevant to the decision. There was no difference in the number of hypotheses generated or in diagnostic accuracy on the variable of years of study. This study demonstrates that nursing students effectively acquire and use domain-specific knowledge over the course of a program of study.

The literature pertaining to decision-making summarized here has important implications for a study addressing the construct validity of a clinical judgment rubric. The literature does support the idea the process involved in decision-making will reveal the developmental level of the nurse. The ability to interpret and make sense of data, set priorities and use the response of the patient to confirm or disconfirm judgments can be indicative of the developmental level of the nurse. The literature also suggests situational complexity will be an important aspect in identifying clinical judgment ability. In order to reveal the ability to handle disconfirming cues and generate alternative hypothesis a situation must be sufficiently complex.

The Literature Pertaining to Reflecting

The theoretical link between reflection and learning is clear and emphatic. Over the course of the 20th century, reflection has become an increasingly central focus for framing how understanding develops. Dewey (1933) identified reflection as “active, persistent, and careful consideration of any belief or supposed form of knowledge in the light of the grounds that support it and the further conclusions to which it tends” (p.6). Dewey (1933) also identified uncertainty as one of the important central elements to reflective thinking. The thinking that occurs in situations of uncertainty and instability forms the basis for Schon’s (1983) critique of positivism and advocacy for the development of reflective practice. Schon discusses the inadequacy of a general scientific theory to explain divergent data that is present in unique and complex situations. The drive to address the discrepant data that does not fit what is generally known is the drive that prompts the practitioner to reflect-in-action and reflect-on-action, which is how Schon differentiates current versus retrospective reflective thought. Mezirow (1994) also identifies uncertainty as a trigger for reflection. Mezirow’s transformation theory (1994) describes learning as a social process of constructing or revising the meaning of experience as a guide to action. Reflection is triggered when old patterns of thinking or old beliefs no longer function to explain the new experiences. Mezirow’s (2000) work is primarily focused on the adult learner and explores how to facilitate critical reflection to achieve transformative learning, which is described as a qualitatively substantive change in thought or action. Boyd and Fales (1983) studied the process of reflection and reached similar conclusion but note that not every new understanding is acted upon. Critical reflection is essential to learning but critical reflection is not necessarily transformative.

King and Kitchener (1994) examine reflection from the perspective of developing judgment and thus focus on the link between reflection and critical thinking. They critique the view of critical thinking as a set of skills or dispositions along with the educational assumption that if the skills and dispositions are the focus of teaching and learning the student will become a critical thinker. King and Kitchener (1994) propose that it is the epistemological assumptions of the learner that differentiates the level of critical thinking. They present a model of reflective judgment identifying seven stages of thinking. Each stage differentiates a specific view of knowledge and concept of justification. The stage one pre-reflective thinker views knowledge as absolute and concrete obtained with certainty by direct observation. The stage one thinker sees no need for a concept of justification because there is an “absolute correspondence between what is believed to be true and what is true” (p14). At the other extreme, stage seven reflective thinking views knowledge as the outcome of the inquiry into complex problems in which the outcome is evaluated against current evidence and reevaluated when new evidence becomes available. Justification at stage seven examines beliefs that are interpreted against a variety of elements such as the weight and value of the evidence, the risk of error, and the possibility of alternative conclusions. The power of the King and Kitchener (1994) model of reflective judgment lies in the fact that no matter what trigger event sets off the subject of the reflection, an examination of how the issues are conceptualized has the potential to reveal the level and depth of the learners' critical thinking. While King and Kitchener (1994) have not adapted their Reflective Judgment Interview to the domain of nursing, they have conducted extensive reliability and validity testing of their reflective judgment model and instrument. They have concluded that the educational setting does seem to facilitate development of reflective judgment.

Metacognition refers to the ability of the individual to predict their performance on a particular task from a self-assessment of their level of mastery and understanding (Bransford, 2000). While the theoretical connection between reflection and learning and reflection and metacognition is strong, making a practical connection between the educational methods used to elicit reflection and the measurable outcome of reflection is less strong. There are barriers to the evaluation of reflection. A primary objection reported in the literature is the validity of the student output. Some concerns identified conclude that writing for a grade has the potential to impact content and writing ability could obscure the level of reflection (Brown & Sorrell, 1993; Wagner, 1999; Wallace, 1996). There is also the concern that faculty would be unable to maintain consistent expectations (Kennison & Misselwitz, 2002). A second aspect of the validity of a student's reflections relates to the student vulnerability during the process. Students who reflect on situations of uncertainty or complexity may reveal themselves in less than an optimal light, faculty may carry forward impressions of student capability that impact the student's evaluation (Pierson, 1998; Platzer, Blake & Ashford, 2000; Steward & Richardson, 2000).

Despite the barriers related to the evaluation of reflection, there is a body of literature supporting the impact of education on the development of reflection. Qualitative research supports the finding that reflection prompts students to think critically about their experiences. Fonteyn and Cahill (1998) integrated a learning strategy of such reflective prompts and found that as students progressed over the junior year of a baccalaureate program their repertoire of thinking strategies increased. In their studies of reflection Sedlak (1997) and Smith (1998) separately found beginning nursing students to be focused on their emotional response to nursing being particularly stressed by the responsibility of initially caring for live, sick individuals. Sedlak (1997) recommended that faculty establish a caring environment with active learning experiences to support student patience in the acquisition of new complex understandings. In addition to the emotional and professional development focus to initial learning, Smith (1998) found evidence that students moved from a position of acceptance of information to one of active critique of evidence and assumptions and stated reflection supports integration of practice experiences with academic knowledge.

The ability of faculty to differentiate the level of reflective capacity with students was the focus of three studies. Wong, Kember, Chung and Yan (1995) used Mezirow's conceptual framework to differentially define levels of reflection as non-reflective, reflective and critically reflective. Pee, Woodman, Fry and Davenport (2002) used a structured framework of cue questions to prompt student reflection based on Boud, Keogh and Walker's (1995) conceptual model. Boenink, Oderwald, De Jonge, Van Tilburg and Smal (2004) had medical students view four vignettes and write their reflections. All three studies found that faculty could use a specific episode of student reflection and differentiate level of ability. This summary of studies demonstrates reflection can be assessed where there is a clear conceptual framework for the different levels of reflection and a clear description of each category. From a validity standpoint, the range of reflective ability identified by the faculty supports its use in the evaluation process.

Two studies examined the development of reflection over time. Murphy (2004) examined the effect of an intervention of instruction and practice with focused reflection and articulation on student nurses. Students randomly assigned to the treatment group were provided with instruction on focused reflection and articulation to facilitate their ability to connect clinical experiences to theory content; this method was practiced and reinforced over two semesters of clinical practicum. Practical knowledge was assessed by a researcher designed instrument used to evaluate the student's assessment and analytical ability. Domain-specific knowledge was assessed by traditional multiple choice examinations. While there was no difference between the treatment and control groups on their composite scores, there was a significant difference, accounting for 29% of the variance, between the two groups on their level of practical clinical knowledge. This finding suggests that instruction, role modeling and student practice with reflection will promote acquisition of clinical reasoning.

Kuiper (2002) studied the change in the cognitive and metacognitive processes of thirty-two new graduate nurses over the course of an eight week preceptorship. A verbal protocol analysis was used to examine the data present in 239 weekly journals. The content of the journals was written in response to prompts based on Kuiper's self-regulated learning theory. Kuiper (2002) found the metacognitive processes of new graduate nurses developed from lower level thinking statements at the beginning of the preceptorship to the ability to connect critical thinking skills with clinical reasoning strategies such as comparison and pattern matching toward the end of the preceptorship. This study demonstrates that structured reflection does facilitate the development of clinical reasoning.

This literature review was focused narrowly to specifically examine the influence of education on the development of reflective thinking ability. However, the information summarized here is consistent with two broader reviews of the reflection literature (Kuiper & Pesut, 2004; Ruth-Sahd, 2003). In summary, the literature supports the many benefits that accompany effective reflection. Reflection promotes integration of knowledge and the knowledge can be the theoretical, practical, emotional or professional. The literature suggests that the experience of a triggering event can promote reflection and that an analysis of the learner's reflective processes can differentially reveal levels of reflective ability. This all supports the use of the high-fidelity simulation setting as a place to conduct such as assessment. The control possible within simulation will allow triggering events to be planned and scheduled and the live action component of the simulation would reveal the student's reflection-in-action capabilities while the debriefing component that follows would engage the student's reflection-on-action.

A Conceptual Model for Learning in High-Fidelity Simulation

Conceptual models relating to simulation learning are rare. In fact, in nursing literature only one model of simulation learning was found (Jeffries, 2005). The Jeffries (2005) model examines the relationships among the various educational practices pertinent to simulation and how they relate to simulation design features and learner outcomes. This model is helpful when designing, implementing and evaluating simulations. However, the model was rejected from use in this study because it was designed to facilitate comparisons between simulation and other educational methods and not designed to explain why an individual learns as a result of participation in a simulation experience. In view of this lack, a conceptual model was developed that examines the relationships between the processes that occur in the high-fidelity learning environment and how these relationships support effective learning.

Learning with understanding occurs within the setting of high-fidelity simulation (HFS) as a result of five interacting processes. A schematic presentation of these processes is represented in Figure 2. The conceptual processes that result in learning are recurring practice, increasing complexity, the high-fidelity nature of the practice experience, feedback, and reflection. The inter-relationships, as well as the theoretical foundations supporting their presence within the model will be examined next.

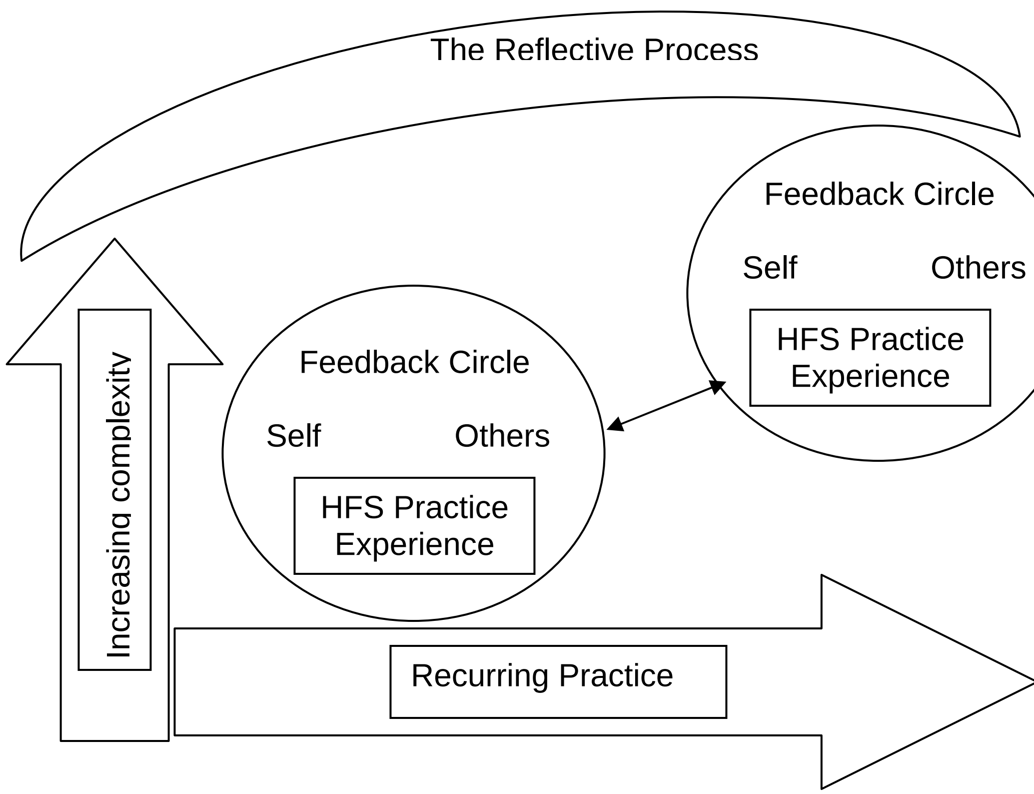


Figure 2. A conceptual model of learning in high-fidelity simulation

Recurring Practice

Recurring practice is possible because of the simulation setting. The experience can be focused on the learner's opportunity to deliberately practice a specific topic at an arbitrarily established time. The recurring practice in HFS is both pedagogically controlled and learner-centered. The topic of learning is identified by the faculty in relation to the learning needs of the student and no longer dependent on the chance occurrence of a similar opportunity in actual practice. The simulated nature of the experience also allows the learner to experience a full range of practice events; since there is no risk to the "life" of the mannequin, the learner is not restricted to experiencing challenges in a subordinate role as they would be should the experience occur in the real-world. The recurring practice available because of HFS allows learners the opportunity to experience and develop practical knowledge and to refine that knowledge in subsequent repeated practice exposures.

Ericsson's (2004) theory of deliberate practice is an important foundation for learning in simulation. The role of deliberate practice in the development of expertise states a key challenge is to avoid the stage of arrested development associated with automaticity. Ericsson describes the process of skill acquisition as one beginning with the practitioner initially focused on the task in order to avoid gross mistakes. As performance becomes smoother and salient mistakes increasingly rare, practitioners no longer need the same level of concentration to perform at an acceptable level. This stage is problematic because as performance becomes increasingly automatic the practitioner loses conscious control and with it the ability to make specific intentional improvements. Deliberate practice to develop expertise can take several forms. The learner can deliberately seek out challenging situations that exceed their current level of performance or the learner can focus on identification and correction of weaker components to improve their ability to anticipate as a means of enhancing performance.

Deliberate practice differs from recurring practice in high-fidelity simulation in two aspects. First, deliberate practice requires metacognition on the part of the learners, while recurring practice alone does not inherently necessitate this. Second, while deliberate practice is deliberate, it is ultimately dependent on chance. The recurring practice in simulation is not dependent on chance for the opportunities to practice. The teaching and learning that occurs in high-fidelity simulation requires both feedback and reflection on recurring practice to promote the development of a deliberate practitioner.

Increasing complexity

Controlling the increasing complexity of the learning environment is possible in the HFS setting. The simulation component of the learning experience allows the educator to adjust the level of learning challenge to appropriately meet the needs of the learners. This means that learning in the setting of HFS can be constructed with the learner building new knowledge on a foundation of existing knowledge through a personal framework of experience, albeit a simulation experience. Gradually increasing the complexity of the simulation learning has the potential to enhance transfer to actual practice as it allows the learner the opportunity to reify known content and integrate new theoretical content into their practice.

Vygotsky's (1978) zone of proximal development supports the conceptual design of increasing complexity to learning within the simulation setting. The zone of proximal development is defined as "the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers" (Vygotsky, 1978, p. 86). As discussed by Chaiklin (2003) the purpose of the zone of proximal development is to identify the types of maturing psychological functions and associated social interactions that are needed to transition from one developmental stage to the next. To place this within a nursing context, the purpose of identifying a student's zone of proximal development, then gradually increasing the complexity of the simulation setting would be to facilitate their progress in the development of professional nursing expertise.

Vygotsky's zone of proximal development carries with it several implications for teaching and learning. The first is collaboration. What can be achieved with the assistance of others is more indicative of potential performance than what can be accomplished alone. The second is to re-conceptualize the meaning of imitation. To Vygotsky a child cannot imitate anything; imitation "presupposes some understanding of the structural relations in a problem that is being solved" (Chailkin, 2003, p. 51). The third is this zone emphasizes readiness at the upper boundaries of competence and the upper boundary is constantly changing with the learner's increasing level of competence (Bransford, 2000). Scaffolding strategies function to assist the performance of the learner in order to promote their transition to unassisted learning (Saunders & Welk, 2005). When considering the complexity of nursing practice and the level of responsibility required of professional nurses, one view of high-fidelity simulation is as a scaffolding technique to support the transition of nurses to the next developmental level.

Feedback Circle

At the most basic level, feedback is information about what happened. Within the setting of high-fidelity simulation, feedback comes from a variety of sources. The situation itself, as a representation of an actual clinical problem provides feedback to the participants regarding the effectiveness of their responses. Participants capable of receiving feedback during active practice are experiencing what Schon (1983) would call reflection-in-action. Other sources of feedback include other individuals involved in the simulation and those watching the simulation and the individual receiving feedback can themselves be a source of information. These sources provide information after the event and thus Schon (1983) would label this type of feedback part of reflection-on-action.

Debriefing a simulation is the process of facilitating reflection on the actions that have just occurred. The process of debriefing the key learning experiences of the simulation presents faculty with the opportunity to discuss the students' background knowledge and perceptions and how that knowledge impacted the experience. The feedback provided through debriefing has been identified as the most critical component of simulation learning (Issenberg, McGaghie, Petrusa, Gorden & Scalese, 2005). The debriefing discussion typically encompasses the "4 E's": events, emotions, empathy and explanation (Mort & Donahue, 2004) as the focus for feedback. It is through feedback during debriefing that learners are able to constructively build on their knowledge base and gain confidence in their abilities (Jeffries, 2005; Lederman, 1992). The process of reflection and guided analysis with the goal of integrating the new understanding into their knowledge base increases the likelihood of transfer from one learning environment to other situations (Bransford, 2000). The reflection required to provide effective feedback during debriefing works to assist the learners to integrate theory and practice and research and allows them to validate their clinical judgments with an appropriate and effective audience.

The High-Fidelity Simulation Practice Experience

The HFS practice experience is a component in this framework because it is through the practice experience that learning stays learner-centered. It is through the simulation practice that the learner comes to experience their level of understanding and mastery of the essential knowledge needed to function effectively as a professional nurse. The learner brings to the practice experience their existing knowledge and conceptions along with their misconceptions. The simulation experience is structured to provide learners the opportunity to practice the application of a nursing standard of care that is built from a research foundation of evidence. The student's ability to implement that standard of care will reveal their understanding of that standard as well as their misunderstandings. The best feedback, according to Wiggins (1998) "is highly specific, directly revealing or highly descriptive of what actually resulted, clear to the performer and available or offered in terms of specific targets and standards" (p. 46). Thus the best feedback is a performance task in which the learner has to confront the effects of their actions directly (Wiggins, 1998). From this perspective, the HFS practice experience is the best feedback possible to provide a learner; it is knowledge-centered but also provides ongoing formative feedback to facilitate improved knowledge and performance.

The Reflective Process

The reflective process in this framework is depicted as an umbrella concept that arches over all the components and links them together in a formative whole. Reflection is foundational to progress in every aspect of the model. Recurring practice is ineffective without reflection. Practice without feedback is ineffective. Practice alone has been shown to result in performance that is more fluid but without improvement in substance and accuracy (Bjork & Kirkevold, 1999). In other words, without effective feedback, nurses are able to perform faster but continue to make the same mistakes leading to the presumption that performance is improving when in fact it is not. Feedback facilitates the development of student reflective capacity (Juwah, Macfarlane-Dick, Matthew, Nicol, Ross & Smith, 2004). Reflection on feedback is the process that links recurring practice to improvement in performance.

The umbrella of the reflective process is tilted at an angle within the model to represent the fact it, too, is an ability that will improve with ongoing practice. The goal of reflection is for the learner to become metacognitive. Learners need practice self-assessing what they know and what they need to learn in order to effectively provide care. In the setting of HFS this process is iterative, simulation after simulation on a gradually increasing level of complexity. Learners also need the opportunity to reflect on their practice of clinical judgment in a direct and open environment. The setting of HFS provides the opportunity for realistic practice of clinical judgment, and this practice has the potential to facilitate development of competence.

Summary of Literature related to Learning in High-Fidelity Simulation

Nursing Literature

A review of the high-fidelity simulation nursing literature was conducted to identify prior studies relating to clinical judgment performance appraisal in this setting. Inclusion criteria included the use of a high-fidelity patient manikin as a teaching instrument and the examination of clinical judgment, clinical decision-making, clinical problem-solving or clinical competence in the populations of both students and practicing nurses. Studies using standardized patients, computer-assisted simulation, or low-fidelity task trainers were excluded. Primary sources of information were the CINAHL and MEDLINE electronic databases but reference lists were also accessed as sources of material. Finally, a manual search of the tables of contents of nursing education journals of the past five years was conducted. Fourteen studies were found.

The studies reported here represent a broad range of methodological rigor. Two of the studies had a qualitative focus (Lasater, 2005; Henrich, Rule, Grady & Ellis, 2002) while the remaining studies were quantitative. Of the quantitative studies, nine evaluated the intervention of high-fidelity simulation using student self-report measures; of these, five did not report statistical analyses (Childs & Sepples, 2006; Henneman & Cunningham, 2005; Rauen, 2004; Spunt, Roster & Adams, 2004; Vandrey & Whitman, 2001); four studies did report statistical findings (Bearnson & Wiker, 2005; Feingold, Calaluce & Kallen, 2004; Larew, Lessans, Spunt, Foster and Covington, 2007; Long, 2005). Two studies examined the effect of HFS learning through assessment of change in cognitive knowledge (Jeffries & Rizzolo, 2006; Nehring, Ellis & Lashley, 2001) along with student self-report of change. The final study was designed to measure clinical practice (Radhakrishnan, Roche, & Cuningham, 2007). All fourteen studies used hospital-based populations as the setting of the simulation scenarios. Twelve of the fourteen studies focused on nursing students with the remaining two using simulation with novice registered nurses in orientation courses.

Of the fourteen studies, four used the high-fidelity setting to evaluate performance. Of this group of four, two studies reported the development of evaluation instruments to examine the impact of high-fidelity learning on nursing practice and two reported anecdotal comparisons. Only the Lasater work specifically addressed the construct of clinical judgment. Lasater's study had a dual focus; the quantitative aspect examined the development and pilot testing of the clinical judgment rubric while the qualitative aspect examined the effect of the HFS experience on the students' development of clinical judgment. The eleven indicators of clinical judgment developed by Lasater for the rubric emerged from her theoretical understanding of the Tanner model as well as her work in high-fidelity simulation. Lasater both observed student performance in the active care of simulated patients and debriefed students post-simulation to gain their reflections on their actions in order to develop the clinical judgment indicators. Lasater observed 42 students in the development of the rubric and then piloted it on another 11 students to refine the instrument.

The second study to use an instrument to evaluation the influence of simulation on nursing clinical practice was conducted by Radhakrishnan, Roche and Cunningham (2007). While this study did not identify a conceptual framework, review of the Clinical Simulation Evaluation Tool developed by the researchers demonstrates an underlying structure of nursing process and critical thinking. The study took 12 senior nursing students in their capstone course and gave the intervention group of six students two one-hour practice sessions in simulation then compared the performance of all twelve on one simulation requiring the care of two patients using the Clinical Simulation Evaluation Tool. The simulation intervention group differed from the control group on the categories of safely identifying their patients and on obtaining vital signs from their patients. There was no difference between the two groups in their abilities to conduct a focused assessment, develop interventions, delegate or communicate. The limitations of the Clinical Simulation Evaluation Tool relate equally to its use of the nursing process and critical thinking as the guiding themes. Nursing literature does not support either as being indicative of actual nursing clinical practice. However, this study does demonstrate the interest of the nursing community in the development of an effective performance evaluation instrument.

The remaining two studies (Larew, Lessans, Spunt, Foster & Covington, 2007; Long, 2005) report anecdotal performance comparisons between experts and novice nurses. The study by Long (2005) used a mock code simulation with both students and a group of physicians and staff nurses. Long reported that the nursing students fit the novice to advanced beginner stage. In a situation of stress and complexity, nursing students struggled in the absence of clear rules and guidelines while the more expert physicians and nurses were able to establish relevancy of cues and set priorities more easily. The work of Larew, Lessans, Spunt, Foster and Covington (2007) describes a method of simulation design based on Benner’s theory. Their scenarios were piloted with a group of experienced critical care nurses before being used with students. Theses researchers note that the experienced nurses were able to identify more patient problems from vague cueing while nursing students required more specific cueing and focused only on the more obvious problems.

The limitations in these four studies demonstrate the lack of strong research specifically directed at evaluating the impact of high-fidelity simulation education on nursing performance. This lack represents a clear gap in nursing knowledge. However, the larger question concerning the appropriateness of the high-fidelity simulation setting as a site for performance evaluation remains. The qualitative aspect of Lasater's study identified several themes that emerged from her work in simulation addressing how students' experiences affected development of clinical judgment. If the remaining twelve studies are examined from this lens, the support for the use of the high-fidelity setting will become apparent.

The first theme identified as strength of simulation learning by Lasater relates to the multiple benefits associated with active learning. Lasater reported that students viewed simulation as an "integrator of learning" (2007, p272) meaning that it brought together foundational theory, psychomotor skills practice and prior clinical work. The fidelity of the simulation experience required students to think about the actions taken with their patients. This theme of integration of learning was present in the findings of four other studies. Feingold, Calaluce and Kallen (2004) reported that 92.3% of their students valued their simulation experience for the manner in which it reinforced their course work, and prompted decision-making and skills practice. The idea that simulation helps students bring together cognitive and psychomotor skills in the application of thinking and decision-making about a specific patient situation was also reported by Henrichs, Rule, Grady and Ellis (2002) Long, (2005) and Vandrey and Whitman (2001).

Another strength of active learning identified by Lasater (2007) was that simulation required students' to anticipate patient responses. Other researchers also identified the value of learning to anticipate. Long (2005) reported that students wanted more information about their patients so they could be more prepared suggesting their experience in simulation taught them the value of anticipation as well as the importance of key types of information. Vandrey and Whitman (2001) discussed simulation practice as encouraging students to assess the continually changing condition of their patient. Bearnson and Wiker (2005) discuss how students learned the importance of thorough assessment and recognizing abnormal findings in a terminology closely reflecting Lasater's discussion of the clinical judgment abilities of noticing and interpreting.

An additional strength of active learning commonly identified was situational safety for both learner and patient. As Lasater (2007) noted, in simulation, experimentation and failure hold no risk for the patient. The opportunity to think though new ways of responding without jeopardizing patients encourages students to expand to higher levels of competency (Rauen, 2004; Vandrey and Whitman, 2001). A hallmark of the novice nurse is a reluctance to implement responses other than those known to be effective while an experienced nurse view known interventions more as a starting point for patient specific responses. The safe environment of simulation presents students with the opportunity to practice new ways of intervening without increasing risk, thus encouraging their development of expertise.

Three studies examined the theme of active learning in simulation through assessment of acquisition of cognitive knowledge with mixed results. Bearson and Wiker (2005) through a student self-report measure identify that students strongly perceive an increase in their knowledge of pharmacological side effects and emerge from simulation practice as strongly confidence of their skill in medication administration. Nehring, Ellis and Lashley used a design series of lecture, pre-test, simulation practice to post-test one and post-test two, and found students gained significant cognitive knowledge from their work in simulation and this knowledge level was retained. Jeffries and Rizzolo (2006) used a design of lecture followed by three types of simulation practice, paper-and-pencil, static manikin and high-fidelity manikin, then compared cognitive knowledge levels and found no difference in cognitive gains across the three groups. These mixed findings underscore the point that active practice in simulation is a performance demonstration and as Jeffries and Rizzolo (2006) point out, cognitive change is an insufficient assessment.

A second theme identified by Lasater (2007) as impacting the students HFS experience and development of clinical judgment was the paradox of anxious and stupid feeling yet increased awareness of learning. This paradox reflects students' statements of learning the most when they performed poorly or were in situations of stress. This theme is also present in the findings of Childes and Sepples (2006) Long (2005) Rauen (2004) and Vandrey and Whitman (2001). Henrichs, Rule, Grady and Ellis (2002) reported that confidence increased when students became familiar with the setting and were able to demonstrate management of the patient situation but that anxiety was a component in all simulations and with some students this anxiety did not decrease with ongoing practice in the simulation setting. Henneman and Cunningham (2005) reported that allowing student to overcome their anxiety about working with a manikin and in an unfamiliar environment was an essential early stage for students to truly engage in the active learning process. Thus the emotional influence on learning and assessment of the level of student anxiety is one of the keys to the process of engagement in active learning. This reinforces the need for accurately identifying the level of simulation complexity to facilitate matching the level of challenge to a point just ahead of the level of student ability.

A third theme brought forward by Lasater (2007) was metacognition as an important experience that occurs with simulation learning and facilitates student clinical judgment development. Lasater identified the debriefing phase of simulation as time when students are most able to reflect on their experiences and integrate new ways of thinking. Other researchers have addressed the concept of metacognition in alternative terms but the underlying basis of providing students with a specific episode and encouraging self-critique is present in the work of Henrichs, Rule, Grady and Ellis (2002) Long (2005) Rauen (2004) and Spunt, Foster and Adams (2004). Lasater notes (2007) that metacognition is helped by the presence of a clear standard for evaluation. For students and faculty to most effectively analyze behavior, a clear standard of behavior is important. The clinical judgment rubric represents such a clear standard of performance this supports the priority of the construct validity work.

A fourth theme Lasater (2007) found to be common in simulation learning and supportive of student development of clinical judgment was feedback. The importance of feedback from the simulation patient during the active phase was reported by Vandrey and Whitman (2001). Childs and Sepples (2006) along with Jeffries and Rizzolo (2006) reported students felt feedback was one of the most important features of simulation learning. Lasater (2007) reported students' desire for deeper and more meaningful feedback, addressing the untoward patient outcomes possible from their actions in simulation. This feeling of concern for their patients, of wanting to know, of wanting to practice until the active performance is right was also found in the work of Henneman and Cunningham (2005) and Long (2005).

Summary of Nursing Findings and Identification of Gaps. At this point in time, the research addressing the development of clinical judgment in the setting of high-fidelity simulation is both limited and lacking rigor. Of the fourteen studies, the most consistent measure of effectiveness reported was student satisfaction. Studies are needed that link simulation learning to the theoretical foundations in both learning and clinical judgment. However, there are features that Lasater (2007) identified as present in the simulation learning experience that do support the development of clinical judgment – active learning, increased learning in paradoxical situations, metacognition and feedback – and these features were consistently present and considered important by other researchers. This review of literature does support the use of HFS as a location for the development of clinical judgment. It is logical that the setting would support a performance appraisal of the same construct. It is also clear from this review of literature that a clinical judgment performance appraisal instrument is needed. Only the Lasater dissertation (2005) presented an instrument designed to assessment clinical judgment performance. The nursing profession needs a valid performance outcome measure such as the Lasater Clinical Judgment Rubric therefore a construct validity examination of this instrument is crucial.

Medical Literature

Human patient simulators have been in use in the field of medicine since the late 1960's (Gaba, 2004). They were primarily developed to facilitate the training of anesthesiologists so the literature in that specialty is relatively rich in research concerning the efficacy of HFS in teaching, training, and performance assessment. This body of literature is of value to review with a focus on determining if the educational technique of simulation is effective in differentiating among known groups representing the varying levels of ability in the specialty of anesthesiology. It is also of interest to review to review this body of literature to determine if an assessment instrument similar to the LCJR was used effectively.

The medical anesthesiology research is consistent in demonstrating that educational level and years of practice are significant factors in the level of expertise that can be demonstrated in the HFS setting. Devitt and colleagues examined this issue in two studies (1998; 2001) and Murray and colleagues conducted three studies (2002; 2003; 2004). All five studies found that increased education and increased clinical experience improved performance on simulator based evaluation. However, other researchers have found conflicting evidence. Some researchers have found highly experienced practitioners to make significant errors in treatment (Byrne & Jones, 1997; Gaba, 1998) and have also found inexperienced junior medical students to perform well. Given these inconsistencies, it is worthwhile to examine the evidence in the Devitt and Murray studies; in particular to determine the assessment methods used to evaluate performance in simulation and the reliability of their findings.

Devitt and colleagues studied performance in HFS by developing case scenarios with specific embedded problems. They developed a technical scoring system to evaluate performance with no response to the problem scored as a 0, a compensating intervention that reflected an action taken to correct the perceived abnormality was scored as a 1, and a corrective treatment, which represented the definitive management of the presenting problem, was scored as a 2. Gaba (1998) would define this method as an assessment of technical performance since the focus was the adequacy of the actions taken from a medical and technical view. In the Devitt and colleagues first study (1997) they tested the inter-rater reliability of this method on two scenarios and found it excellent at 0.96. The next two studies focused on construct validity to determine if the rating system could discriminate between levels of ability. In the 1998 study participants were anesthesiology residents and the university based anesthesiology faculty. In 2001 the participant base expanded from 4th year medical students, to residents to three levels of practicing physicians. In both of these studies they did find evidence of construct validity, as the evaluation system used in HFS was able to discriminate between individuals based on training and clinical experience. The difficulty Devitt and colleagues experienced in these two studies was internal consistency. In the 1998 study the final internal consistency score was 0.66 and in the 2001 study the final score was 0.69, with the desired range identified as > 0.60 and <0.90. Internal consistency is a measure of reliability that identifies how consistent the results are on the different components in the checklist. Several possible sources of error explaining this low internal consistency were identified. The first source could the low number of items on the checklist as a statistical source of decreased reliability. The second source could be variability in rating, in the 1998 and 2001 studies Devitt and colleagues did not calculate the inter-rater reliability having previously documented it at 0.96. The third source of variability is that subjects could vary their performance from scenario to scenario and item to item and this suggests that content validity of the items in the performance checklist may have lacked rigor. Actually this highlights a disadvantage of using performance checklists. For performance checklists to be an effective tool of measurement the evidence base defining performance needs to be clearly established. Unfortunately, between the current state of medical science and the complexity of the average situation, even in the controlled environment of simulation, because it is high-fidelity, this level of objectivity is rare.

Murray and colleagues took an alternative approach; they used a variety of scoring methods and used generalizability theory to examine the sources of variability. A core component in all three studies was construct validity to determine if the scoring system could discriminate between levels of ability. In 2002, third and fourth year medical students along with residents were compared. In 2003, first year residents were compared to fourth year residents. In 2004 junior residents were compared to senior residents. In all three studies they did find a difference attributable to education and clinical experience. Murray and colleagues developed an analytic scoring system. Analytic scoring is a checklist of important and essential behaviors and actions that could be expected in the course of the simulation; this checklist could be weighted to reflect the relative importance of the various items. They also developed a holistic scoring rubric with three dimensions: thought processes, actions and integration, and overall. In the 2002 study they found that work in the scenario development using content experts and item analysis is foundational to reliability results. They report the inter-rater reliability of four raters at 0.87 and state if care is taken in the development of the scenario and the scoring rubrics, reliable and valid measures of clinical performance can be obtained. Their findings in the 2003 study echo this conclusion but go on to state the most variance in performance was in the person by case component which suggests that multiple scenarios are necessary to accurately evaluate performance. The need for using multiple scenarios would suggest that a simulation complexity rubric would be important in order to identify both an appropriate level and ensure that a range of performance levels are being evaluated. In the 2004 study the generalizability analysis determined that variability of subject by rater was small indicating the criteria were being consistently applied. They also found the rater by case variability was small indicating raters rank ordered the complexity of the cases in a similar fashion and used the rating systems consistently. Again, they found the most variance was subject by case suggesting that reliability was more dependent on the number of scenarios rather than the number of raters. This was also the conclusion of Weller and colleagues (2005) who found the participant accounted for 27% of the variance in the scores and the interaction between the participant and the case accounted for 37% of the variance in scores. They also recommend multiple scenarios be used to reliably rank performance.

There is a third approach to evaluation of clinical performance in simulation. Crisis resource management (CRM) is a view on performance evaluation that grew from research in the aviation field. CRM places equal value on both assessment of technical performance and assessment of behavioral performance. Behavioral performance assessment examines the decision-making and team interaction processes used during the management of a situation (Gaba, Howard, Flanagan, Smith, Fish & Botney, 1998). Gaba and colleagues have identified ten crisis management behaviors: orientation to case, inquiry/assertion, communication, feedback, leadership, group climate, anticipation/planning, workload distribution, vigilance and reevaluation. These behaviors are rated on a 5 point ordinal scale with 1 equal to poor performance and 5 equal to outstanding performance. Two summary behaviors are also scored, one for the primary team leader and one for the overall performance of the team. Technical behaviors were also scored from a checklist of appropriate medical and technical actions pertinent to each scenario.

Gaba and colleagues (1998) found three raters agreed 86% of the time on the presence or absence of a technical action in the cardiac arrest simulation and 83% of the time on the malignant hypertension scenario. They note the utility of reliably scored technical performance to be the identification of people who do not carry out one or more essential actions can clearly be viewed as deficient, whereas those who carry out only those essential actions are minimally acceptable, and those who perform multiple appropriate actions are identifiable as performing well. The inter-rater reliability results with the behavioral scoring showed greater variation. Using a statistically conservative method the IRR ranged from 0.32 to 0.54 for the two overall ratings; using a more liberal statistical method the IRR ranged from 0.6 to 0.93. The difficulty with rating behavior is that over the course of a scenario the behavior can vary markedly. Gaba and colleagues conclude that a perfect scoring system for complex clinical situation will be difficult to achieve but is it likely to require multiple raters assessing both behavioral and technical performance over several episodes of simulation.

Summary of medical findings and identification of gaps. The medical literature, in particular the work of Deavitt and colleagues (1998; 2001), Gaba (1998), and Murray and colleagues (2002; 2003; 2004) supports several conclusions. First, education and practice do influence the level of clinical ability. Second, these differences in ability are apparent in performance assessment conducted in the high-fidelity simulation setting. Third, performance in simulation varies with the individual's knowledge of the content (Morgan, Cleave-Hogg, McIllroy, Devitt, 2002; Byrne & Jones, 1997).

However, the medical simulation literature is less clear regarding scoring systems.

Technical scoring systems, such as performance checklists, identify key actions that are planned to occur during the course of the simulation. The level of action initiated can then be used to evaluate the response of the examinees, typical choices include taking no action, managing symptoms through compensatory actions or responding with a corrective action. Behavioral scoring systems allow researchers to identify selected behaviors and define a range of performance levels. The reliability of behavioral scoring systems has historically been less than the reliability of technical scoring systems.

Both scoring systems have inherent limitations. Technical scoring systems are response based or ‘ends’ oriented. They do not reveal the process the learner used to determine their choice of action. If the evidence base supporting the response is clear then this type of scoring system can be effectively written and will reveal those learners who often, sometimes or never know how to respond to a specific situation. The utility of a technical scoring system in ambiguous situations has not been addressed in the literature.

Behavioral scoring systems are process based or ‘means’ oriented. By defining a range of performance levels for a set of behaviors, researchers are attempting to evaluate how effectively a learner engages in a specific behavior. The effectiveness of engagement in behaviors has been historically difficult to evaluate. However, if the focus of education is to develop learners who can take their body of knowledge and use it to effectively interact in new situations or ambiguous situations, then assessment of essential behaviors is a goal worth pursuing.

Summary of Literature Findings

The ability to make sound clinical judgments is a competency central to the practice of nursing. This review of literature regarding the construct established the clear need for a reliable and valid competency assessment instrument. Historical attempts to measure clinical judgment were found to be insufficient. An inadequate understanding of the clinical judgment process had been a primary problem. The research-based model of clinical judgment (Tanner, 2006) that emerged from an extensive synthesis of research clarified this issue. The Lasater (2005) Clinical Judgment Rubric operationalized the process of clinical judgment by using the dimensions present in Tanner's model and differentiating level of ability on a novice-to-expert type developmental continuum.

When the purpose of an instrument is to determine how much an individual possesses of a particular trait or quality that is presumed to be reflected by performance on the instrument, construct validity becomes a priority (Waltz, Strickland & Lenz, 2005). The construct validity process requires examining the relationship between the operationalized instrument and the theoretical foundations. The theoretical foundations supporting the Tanner model of clinical judgment provide testable assumptions for the examination of the construct validity of the Lasater Clinical Judgment Rubric. To summarize these theoretical foundations, the testable relationships as revealed by the review of literature, include the following:

1) Level of ability in clinical judgment increases with the level of knowledge and experience possessed by the nurse
2) Level of ability in clinical judgment demonstrated by the nurse will vary with the complexity of the situational context

A research design using the known-groups method facilitated the examination of these assumptions. The use of two groups with known differences in level of clinical judgment ability allowed for the examination of the impact of domain-specific knowledge and experience. Locating the study within setting of high-fidelity simulation was useful for multiple reasons. First, the location allowed examination of the second assumption through the control of contextual complexity. Second, simulation allowed for consistency of case presentation across the study participants. Third, the level of situational fidelity provided an opportunity to observe the application of the clinical judgment process under realistic conditions. The methodology presented in chapter three further describes the design features used in this study.

CHAPTER THREE

METHODOLOGY

The primary focus of this study was evaluation of the construct validity of an instrument of purporting to measure clinical judgment, the Lasater Clinical Judgment Rubric (LCJR). It is becoming imperative that nurse educators determine, with some certainty at program completion, that students are competent in their application of clinical judgment. The high-fidelity simulation setting offers unique opportunities for this type of evaluation. Secondary questions that were examined within this study included how the complexity of simulation scenarios affects the demonstration of clinical judgment and how faculty evaluation compared with student self-evaluation on the level of clinical judgment ability demonstrated. This chapter describes the methods developed to address these issues.

Study Design

Known groups' methodology was applied to examine the issue of construct validity (Waltz, Strickland & Lenz, 2005). The hypothesis applied to this cross-section of students was that graduating senior nursing students would demonstrate a significantly higher level of clinical judgment as measured by the LCJR, than end-of year junior nursing students as a result of their increased domain-specific nursing knowledge and amount of clinical experience. In order to draw inferences about the validity of the LCJR, it was important that the groups were equivalent except in the amount of nursing experience. This was important because of the assumption that clinical judgment would improve as a function of clinical nursing experience. Since the groups were not randomized, it was possible there would be differences that could obscure actual differences in clinical judgment ability, such as variations in life experience that might be related to clinical judgment ability. To control for these initial differences, data were collected on a number of variables that might be related to clinical judgment ability, such as prior experience in caregiving and prior experience in the educational setting. Other possible influencing variables included age, gender, GPA on admission to the nursing program and demographic data were collected on these variables as well.

Sample and Setting

Sample

The target population for the study was baccalaureate nursing students. A convenience sample was used. The experimentally accessible population was comprised of those students enrolled in two existing courses at OHSU. The two courses represented natural breakpoints in the two year nursing curriculum, with one course providing the sample pool for the end-of-year junior students and the second course providing the sample pool of graduating seniors. This sampling method limits the generalizability of the study findings.

A high-fidelity simulation learning activity was a required component in both courses. While the simulation learning activity was required, participation in the research study was voluntary. The group of junior students consisted of those enrolled in the level one adult medical-surgical clinical course (n = 27). The group of senior students consisted of those enrolled in the level two medical-surgical course (n = 45). All but one student chose to participate in this research study. No students were lost to the study through attrition for academic reasons.

Sample size. Findings from the Lasater (2005) study were used to estimate effect size in order to complete a power analysis. Lasater's work was conducted with junior level students who scored a mean of 22.98 with a standard deviation of 6.07. A power analysis was conducted for an alpha of .01 and .05. The results of the power analysis indicated that a sample size of at least 16 was needed for an alpha of .01 and a sample size of 11 was needed for an alpha of .05 in order to detect a significant difference between groups if the seniors' scores were in the expected range of 33. A small sample size such as 11-16 would only be acceptable if it was known that the LCJR was a precise measurement instrument and it was known that large differences are present in the two groups being compared. However, at this stage of research, uncertainty exists regarding possible within group differences, the impact of situational complexity on performance and the potential for measurement error. In view of these factors, a sample size of 24 participants for each group was identified as an appropriate balance between power and researcher burden.

The number of junior students enrolled in the pertinent course was 27. One student chose not to participate in the study. Of the remaining 26 students, four were lost secondary to recording problems making the final sample size for the junior group 22 participants. The senior group was larger with 45 students enrolled in the course and all choosing to participate in the study. The data from two students was lost in recording malfunction. Of the remaining 43 students, all of who volunteered to participate, 24 were selected at random for evaluation.

Setting

The Ashland campus of the Oregon Health and Science University (OHSU) School of Nursing was selected as the site for this study. The OHSU School of Nursing offers a fully accredited program of study in the nursing major on four campuses across the state of Oregon. The program consists of two years of upper division course work leading to a Bachelor of Science degree as well as preparing students to become licensed registered nurses in the United States. Throughout the program students register for theory courses that are paired with clinical practica in a variety of sites, thus meeting the mandates of both the national accrediting bodies and the Oregon State Board of Nursing. The design of the nursing curriculum, condensing the domain-specific courses into two years, facilitates the comparison of two levels of student ability within this study design. It was a natural point of comparison to examine the stage of development of the two groups of students.

While simulation learning was occurring on all four campuses of OHSU, the Ashland campus simulation facility was selected as the site for this study. The facility was recently remodeled and represented a state of the art simulation center. A control room contained the computer equipment for managing the patient manikin and recording the audiovisual activities occurring in the theater room. The simulation theater room was the location of the interaction between students and the patient manikin. The facility has the technology available in terms of audio and video mixing to record the simulations and later transfer the information to DVDs. Recording the simulation learning experience enhanced the logistics of scheduling the work of both students and faculty. Viewing the DVDs at a later date allowed the hiring of faculty blind to the level of the student being evaluated and strengthened the study design. Also, viewing DVDs of students who were outside the sphere of influence of the faculty protected the student from any possible carryover influence to a course grade.

The simulation faculty working at the Ashland campus were capable of consistently implementing the simulations required in this study. Both faculty members were experienced educators in the traditional practices of post-clinical conferences and both had attended two courses directed toward developing skills specific for use in simulation. At the initiation point of this study, both faculty members had run more than 375 simulations over the past year in the setting. Both of the simulation faculty members dedicated to teaching in the facility, meaning they were involved in providing simulation learning to all students across all courses. These factors are important to note as they pertain to the ability of the faculty to be consistent in their application of the educational method. While both faculty members were involved in running the active component of the simulation and providing student participants programmed cues, the debriefing component was conducted by only by the primary researcher.

Procedures for the protection of study participants. This study protocol was submitted to the dissertation review committee at OHSU in March 2006. Subsequent to approval by the committee it was submitted to the Institutional Review Board at OHSU and met the criteria for an expedited review. Following review and approval by both boards, the study commenced spring quarter 2006.

The privacy and confidentiality of the student participants in this study were protected through several mechanisms. First, participation in the study was voluntary. Information was provided about the study, and students were asked to sign an informed consent (Appendix B). As participation in the simulation aspect of the study was a component built into their course work, all students received equal educational benefit. However, only those students who consented had their recordings copied for review by study faculty clinical judgment raters. Second, all student data were assigned a unique identifying code. Data that needed to be linked during study analysis were matched by unique code only. Third, all data pertaining to the study were stored in a locked cabinet and accessible only to individuals working on the study. Fourth, the faculty viewing the DVDs and scoring the student's performance were hired from outside the university and not associated with any of the internal graded student assignments.

Instruments

Clinical judgment is a context specific nursing ability. The level of clinical judgment demonstrated becomes apparent as the student engages in a specific patient situation. The Lasater Clinical Judgment Rubric (LCJR) is an observational tool designed to facilitate faculty assessment of clinical judgment level. However, as it is the combination of student in the situation that reveals level of clinical judgment ability, a second instrument that measured situational complexity was also necessary. The third instrument necessary was a record of student demographic data. The demographic data were used to both identify the students' position in the curriculum and to identify the presence of any pertinent background experience that may impact demonstration of clinical judgment.

Lasater Clinical Judgment Rubric

In this study, the Lasater Clinical Judgment Rubric (LCJR) was the instrument used to measure clinical judgment performance (Appendix A). This rubric identifies four levels of clinical judgment starting with beginning, then advancing to developing, accomplished and exemplary. The LCJR also defines four dimensions based on Tanner's research-based model of clinical judgment. The four dimensions are noticing, interpreting, responding, and reflecting. Each dimension is further subdivided into indicators of clinical judgment. The dimension of noticing is comprised of focused observation, recognizing deviations from expected patterns and information seeking. The dimension of interpreting involves prioritizing data, and making sense of data. The dimension of responding requires a calm, confident manner, clear communication, well-planned interventions/flexibility, and skillful implementation. The dimension of reflecting involves evaluation with self-analysis and commitment to improvement. The LCJR takes all of the eleven indicators in the four dimensions and describes each thoroughly across all four levels of clinical judgment. For example, in the dimension of noticing and under the variable of focused observation there is a comprehensive yet concise description of the beginning nurse's capabilities when conducting a focused observation with a patient. There are equally clear, concise and comprehensive descriptions of focused observation for developing, accomplished and exemplary levels of clinical judgment. This process continues for all the indicators in each dimension of the rubric.

Scoring system. The LCJR is an instrument designed to assess the level of clinical judgment exhibited by an individual. The time periods that provided the data for the LCJR include both the active component of the simulation and the debriefing component. Both components served to provide quality information because both were designed around a framework of the clinical judgment process. To score a student, their behaviors and verbalizations were assessed to determine the level of ability represented. Behaviors and verbalizations indicative of the beginning level were rated at 1, those at the developing level rated at 2, accomplished behaviors rated at 3 and exemplary rated at 4. A copy of the scoring worksheet is included in Appendix A.

The data from the LCJR scoring worksheet can be analyzed in multiple ways. Each of the 11 indicators of clinical judgment defined within the rubric provides information about a specific area of ability. The data can also be grouped by dimension of noticing, interpreting, responding or reflecting. Finally, a total score can be summed and averaged to determine the level of clinical judgment demonstrated by a student within a specific context.

The LCJR is a new instrument, developed and piloted as part of Lasater's dissertation work in 2005. In the initial work, the instrument was used to assess the level of clinical judgment of 26 junior students enrolled in the second clinical course of the curriculum. The mean clinical judgment score for this group of students was 22.98 (SD = 6.07), as Lasater reported raw score totals. Averaging the raw score across the eleven indicators transforms the score into a format that is more closely aligned with the numbers associated with the levels of clinical judgment as listed on the scoring sheet and on the rubric. Thus the raw score of 22.98 transforms to a score of 2.09 and locates the pilot group of students at the low end of the developing level, which coincides with faculty expectations given their stage of domain-specific knowledge and clinical experience.

The Simulation Complexity Rubric

Complexity became a central design issue. From a learning theory standpoint, it was important to scaffold the learning activity to challenge the learner sufficiently while not overwhelming their capacity to acquire new understandings. From a performance assessment standpoint, it was important to provide a spectrum of experiences in order to reveal the full range of the learner's abilities. The influence of complexity is not explicitly addressed in the research-based model of clinical judgment; it is only present in the model as a function of the nurses' prior experience with similar situations. In view of the importance of complexity to both simulation design and clinical judgment performance assessment, part of the work put forward by this study was the development of a simulation complexity rubric (See Appendix C).

The purpose of developing a simulation complexity rubric was to aid in the differentiation of the complexity levels present in any particular simulation scenario. The rubric that was developed defined three levels of complexity, low, medium and high and six dimension of complexity. The six dimensions of complexity include the following:

- Learner Pre-Knowledge of the Situation
- Task/Setting Complexity
- Physiological Complexity
- Psycho/Social/Spiritual Complexity
- Information Source Complexity
- Potential Risk/Danger to SimPatient Complexity

There is support for these dimensions of simulation complexity in the literature. In her description of design features pertinent to simulation, Jeffries (2005) includes information, cues, and complexity. In Lewis' (1997) description of task complexity in decision-making the components identified are the content present in the cues, the alternatives possible and the context of the people and environment. What alters each of these components is the number of cues, their relevance, level of ambiguity, conflict and change required. These components were built into the dimensions and levels of the simulation complexity rubric.

The work of Leenders (1998) provides additional evidence for content validity of the dimensions in the simulation complexity rubric. Leenders differentiated complexity along three dimensions: analytical, conceptual and presentation. These dimensions and levels were incorporated into the simulation complexity rubric. Analytical dimensions were addressed through physiological complexity and potential risk/danger with levels ranging from a single problem with low risk to the SimPatient to highly complex situations where the patient physiology is unstable and there is an immediate risk to the SimPatient. Conceptual dimensions included the complexity of the setting and the grouping of psycho/social/spiritual aspects of complexity. Levels of complexity in these dimensions altered with the number of concepts, personnel or equipment included in the scenario. Finally, the presentation dimensions were addressed in the simulation complexity rubric through the information source complexity and the learner pre-knowledge of the situation. Levels in these areas range according to the amount of the material given to the learner, the organization of the material, and the source and reliability of the material. Evidence of content validity at this initial stage was focused on determining if the levels and dimensions in the rubric accurately represented the concept of complexity as applied to high-fidelity simulation.

Simulation scenario complexity was scored by each of the faculty raters in this study (Appendix D). Each rater independently rated the complexity of each scenario on each of the dimensions of complexity listed in the rubric. In addition to their opinions on each of the dimensions, faculty were asked to identify a global complexity rating for each of the three scenarios. Raters were also asked to note any deficiencies in the dimensions of complexity as well as endorse pertinent complexity constructs. This work by the faculty raters occurred after they had completed their student rating.

Simulation Scenarios

Development process. The scenarios presented in the simulation were designed around the research-based clinical judgment model. Forward progress through the simulation was structured around the process of noticing, interpreting, responding, and reflecting. For example, in the simulation with the learning goal to demonstrate competence in the care of the patient with an acute episode of congestive heart failure, the student was given cues to notice and prioritize. In this case the cues were the patient subjectively complaining of dyspnea and objectively presenting with a respiratory rate of 26 per minute. The student was expected to respond by assessing the patient's respiratory status with a variety of assessment techniques, and then interpret the data in order to respond appropriately. Depending on the learning level of the student the response may be as simple as repositioning the patient with the head of the bed elevated or as complex as needing to report the interpretation of their findings to the patient's physician to obtain new therapeutic orders. The context was realistic. If the student did not notice the initial set of cues, the patient's condition was altered to bring forward a new set of cues for the student to notice and interpret. If the response was inappropriate, the patient's condition altered to reflect this as well. If the student did respond appropriately but did not continue to reassess, re-evaluate the situation, this also caused the simulation to alter. The simulation was designed to engage the student in the active practice of clinical judgment. The simulation did not present problems and cues to notice in isolation but rather layers of problems and health alterations were built into each simulation. How the student responded to the cues and prioritized the problems revealed their level of clinical judgment. Framing scenario development and progress on the model of clinical judgment thus facilitated assessment of student performance on the construct. This interactive and emerging process continued for the 15 minutes of time allotted to the active component of each simulation.

This study used three scenarios with each student participant required to complete the set. The content contained within each scenario reflected adult health alterations that occur commonly in nursing. Using common health alterations in this study served two purposes. First, assessment of performance on common problems reflected the level of student attainment of competencies essential to nursing. Second, the LCJR is designed to assess the students' ability to demonstrate clinical judgment. It was important to avoid the possibility of testing idiosyncratic knowledge of a specific protocol. The three health alterations identified and developed for use in this study included management of heart failure, management of chronic obstructive pulmonary disease, and management of diabetes mellitus (Appendix E). These health alterations were identified as priority areas for health care competence by the Institute of Medicine (2003).

While the three scenarios were developed around health alterations common to nursing, they also reflected a range of complexity. Presenting the full range of complexity was important in order to avoid a type II error. If the scenario presented was too simple, then all students would do well. Conversely, if the scenario presented was too difficult, then all students would perform poorly. The situation of either a floor or ceiling effect would result in finding no difference between the two groups when a difference was actually present. Presenting students with a range of complexity was planned to mitigate this issue.

Student preparation. One week prior to the start of the evaluation process, all students were provided with preparatory information (Appendix E). For all three patients, a history and physical examination was developed along with a set of physician orders for care. With the active simulation time working with the patient in the theater limited to 15 minutes, it was important to provide the students with a focus for their expectations and a direction for their initial noticing. These students had never ‘met’ any of the simulation patients in any prior learning activity so they had no prior knowledge of the patient’s personal values or typical pattern of responses, other then the data provided in the preparatory information. This patient specific background information supplemented the theoretical and practical knowledge the student brought to the encounter. Theoretically the background data would have deeper meaning for the senior students then the junior students, and this deeper understanding would be visible in their level of clinical judgment demonstrated.

Scenario Debriefing. Debriefing a simulation is the process of facilitating reflection on the actions that have just occurred. The process of debriefing the key learning experiences of the simulation presents faculty with the opportunity to discuss the students’ background knowledge and perceptions and how that knowledge impacted the experience. Thus it provides information that is not only indicative of the students’ reflective ability as one component of clinical judgment but also but is also an opportunity to gain insight into the student's thought processes in other areas.

The amount of time allocated to debriefing was 15 minutes, which was the same amount of time allocated to the active simulation component of the performance assessment. To facilitate the faculty raters' consistent assessment of students, rating guides were developed for each of the three simulation scenarios (See Appendix F). The questions built into each of the debriefing guides also followed the research-based model of clinical judgment, starting with questions relating to what the student noticed during their initial interactions with the patient, how they interpreted their findings, then moving into questions concerning how they chose to respond and what factors impacted their priorities for the care provided to the patient. Finally, students were given an open opportunity to reflect on their experience.

Student Demographic Data Sheet.

Demographic data were collected on each study participant (Appendix G). As the background that nurses bring to the setting influences their clinical judgment ability, information was collected identifying the participant's age, gender, amount of prior experience in care giving and their previous experience in learning. The purpose of the student demographic data sheet was to determine the equivalence of the two student groups.

Procedures

Inter-rater agreement testing

Four raters were necessary to complete the workload associated with the volume of clinical judgment rating of students required in this study. The selection criteria for the four raters were specific. All raters possessed a master's degree in nursing, worked full-time in nursing education, and had experience with both the theoretical and practice aspects of simulation and a working knowledge of the model of clinical judgment.

The initial training to develop inter-rater reliability on the LCJR began with a six-hour seminar. The seminar served several purposes. First, it provided all participating faculty raters with a similar baseline understanding of clinical judgment theory and the research-based model of clinical judgment. Second, it served as a forum to bring the raters in contact with the developer of the clinical judgment rubric, Dr. K. Lasater, who discussed the process of the development of the rubric and how the four levels of clinical judgment were differentiated. Third, the seminar provided the opportunity for the raters to engage in the active practice of the application of the clinical judgment rubric. The focus of seminar was to engage the raters in dialogue regarding the rubric. The end goal was to attain group consensus on the interpretation of the meaning of the rubric levels and their application to student performance evaluation.

Three rounds of active practice were completed in this initial training session. The process of active practice rating consisted of the group of faculty raters watching a DVD of a student participant responding to a high-fidelity simulation clinical practice session and its associated debriefing. At the conclusion of the DVD viewing and after the raters had independently completed their clinical judgment scoring, they reported their results. Then raters were encouraged to discuss what they observed in the student performance and how they interpreted the actions in relationship to the clinical judgment rubric. The ensuing discussion was vigorous across all components of clinical judgment; noticing, interpreting, responding and reflecting.

Each of the rounds of rating practice addressed a different simulation scenario, so by the end of the training session, the raters had been exposed to each of the three simulation practice cases used in this study. Round one of practice presented the raters with the low complexity simulation but the student performance emerged as in the middle levels of the rubric. This was problematic for the raters as it required the performance be differentiated from the level below as well as the level above where the student was currently. Level of agreement between raters and between raters and Dr. Lasater differed markedly, from zero to 91%. Extensive discussion among the faculty raters addressed both what was observed as important in the student's performance and how these behaviors were translated into the language of the rubric.

Round two presented faculty raters with the medium complexity simulation. The student performance emerged as at the beginning level. This was easier for faculty to rate, as they only had to differentiate this beginning student from the one level above. Faculty level of agreement improved to 82% to 91%.

Round three of active practice presented raters with the high complexity simulation. The student level of performance that emerged placed her in the mid range of the rubric only this time at the accomplished level. To assess this student's performance faculty had to differentiate it from the developing level below and the exemplary level above. This particular student provided faculty with the opportunity to struggle on the upper ends of the rubric, differentiating the accomplished nurse from the exemplary. As a result, levels of agreement fell, ranging from 55% to73%

At the conclusion of this initial training session, all faculty raters were dissatisfied with their level of agreements and their understanding of the application of the rubric.

The faculty were in agreement in requesting additional time with Dr. Lasater in order to acquire a better understanding of the language of her rubric. In fact, they did not obtain the levels of agreement designed in this study, which were set at 100% agreement regarding level of student (beginning, developing, accomplished or exemplary) and greater than 80% agreement on the individual variables. At this stage one rater withdrew from the study for personal reasons. After discussion of workload with the remaining three raters, it was decided to look for a replacement rater.

Scheduling conflicts and geographic dispersion precluded the group from meeting again in a face-to-face manner. In addition, the need to bring in and train a replacement rater necessitated development of a modular training method. Each training module consisted of a DVD of a student performance in simulation with the associated debriefing. For each of these DVDs, a clinical judgment scoring guide was developed in conjunction with Dr. Lasater. The scoring guide identified the clinical judgment level of the student on each of the variables and provided a brief annotation regarding why the level was the best fit for the student. These modules allowed faculty to view a training DVD, complete their evaluation and submit their results via email and then receive feedback by return email. In this manner faculty were able to compare their assessment with Dr. Lasater's assessments of the student.

Modules were developed from the initial training session DVDs presenting the low, medium and high complexity simulation scenarios. For the supplemental training, it was decided to utilize the same low complexity simulation scenario and vary the student participant. Theoretically, keeping the situation constant and changing the student should have facilitated faculty understanding of the rubric across all the levels of clinical judgment, which should then transfer to other settings. Faculty were asked to continue supplemental training until they attained a greater than 90% level of agreement with the standard set in the module.

The flexibility of the modular method was effective in moving faculty raters forward in their understanding and application of the rubric. The number of supplemental cycles of training varied across all raters from one to four cycles necessary to attain a greater than 90% level of agreement. All raters did attain this level prior to initiating the rating of the study participants.

Inter-rater level of agreement was assessed systematically throughout the course of the study. Overlap DVDs between pairs of raters were scheduled at the fourth, eighth and thirteenth rounds of faculty rating. Level of agreement between the pairs of raters was calculated by determining how often the pair of raters agreed or disagreed on the score for the student over the 11 indicators of clinical judgment in the rubric. Level of agreement was calculated for each of the three simulations. To obtain a mean level of agreement between the pair of raters for each of the overlap rounds, the levels of agreement calculated for each of the three simulations were averaged. While the comparison ratings between raters three and four at round four of rating started out high, the remaining inter-rater reliability in this study was inconsistent (see Table 1).

Table 1.

Longitudinal Inter-rater Levels of Agreement

Round Four		Round Eight		Round Thirteen	
Rater Pair	% Agreement	Rater Pair	% Agreement	Rater Pair	% Agreement
R1 to R2	24	R1 to R3	24	R1 to R4	9
R3 to R4	88	R2 to R4	64	R2 to R3	33

The calculation of inter-rater reliability using level of agreement on each of the 11 indicators of clinical judgment was founded on the assumption that each of the indicators is reasonably independent. Variables nested in sets, such as the indicators nested in the LCJR violate this assumption of independence (Downey, 2004). In this situation the unit of reliability analysis is the case. In this study design, the raters were provided data from three cases, or simulation scenarios, and no other information. To determine the level of agreement between raters in consideration of the nested nature of the clinical judgment indicators, the standard for the level of agreement between raters was expanded by one point. Thus, faculty ratings of performance that differed by one level or less were considered as equal and ratings that differed by two levels or more were considered unequal. Expanding the standard of agreement in this fashion revealed how closely the raters agreed on the level of student performance within the framework of the nested case. The data demonstrate improved levels of inter-rater reliability (see Table 2). The pattern of data suggests the issue with inter-rater reliability was not one of construct under-representation, as across the three cases the raters had enough data to consistently place student performance within the range of two levels. The inter-rater data analysis suggests the threat to reliability was construct-irrelevant variance, which is amenable to improvement with rater training.

Table 2.

Longitudinal Inter-rater Levels of Agreement Using Expanded Standard

Round Four		Round Eight		Round Thirteen	
Rater Pair	% Agreement	Rater Pair	% Agreement	Rater Pair	% Agreement
R1 to R2	73	R1 to R3	91	R1 to R4	85
R3 to R4	100	R2 to R4	100	R2 to R3	57

Clinical Judgment Data Collection

Audiovisual recording of the student's performance in the active component of the simulation scenario as well as the debriefing process was the actual data used to determine the student's level of clinical judgment. Each study participant worked through all three simulations, the low, medium and high complexity scenarios. The active portion was 15 minutes and 15 minutes was allotted to the debriefing portion. Thus the collections of clinical judgment data required 90 minutes for each study participant and in one day a maximum of six participants were scheduled. As completion of the three simulation scenarios was a required component of the course work for both groups, all the students rotated through this process. Each student spent 1½ hours in the simulation theater, involved in active simulation and debriefing. At the conclusion of each individual's debriefing, the student was provided with a copy of the LCJR and asked to self-assess their performance from that day.

Scheduling logistics. A schedule was developed for the days the simulation facility was dedicated to the research simulations. Students were asked to sign up for appointment times and reminded that in addition to the 1 ½ hour of theater time, they would need to allow for time to complete their clinical judgment self-evaluation. A daily log sheet was designed and used to track which students had completed their simulations, their self-assessment and received their certificate of thanks with a $5 coffee card attached. Only two students rescheduled their appointments and only three students needed email reminders to schedule their appointments.

Data analysis procedures. Data were analyzed using SPSS software, version 13.0. All data were entered into the program by the researcher. Accuracy of data entry was confirmed manually, using a two-person reader-to-reviewer technique. A codebook was developed that listed and defined each of the variables entered into the SPSS database.

Data analysis began with an examination of the demographic equivalence of the two student groups. The continuous variables of age, GPA on admission to nursing school, years of prior experience in caregiving and years of post high school education were compared using the independent t-test. The dichotomous variables of gender and previous educational degree were compared using a chi-square test.

In addition data were collected to determine if the factors of time of day when tested or the order of the testing impacted the results in any systematic manner. Regression analysis was applied to this data as the focus of analysis was a possible relationship between variables, not a relationship between groups of students. A regression analysis would determine if a construct-irrelevant factor, such as day of the week or time of day, contributed variance to the findings in the LCJR scores.

The data from the LCJR was analyzed three ways. Variables were created for each of the 11 indicators of clinical judgment in the rubric. Thus, the first set of analyses examined clinical judgment by each specific indicator. A second set of variables was created in SPSS using the transform – compute function to summarize the data by dimension. This data facilitated clinical judgment comparisons along the dimensions of clinical judgment – noticing, interpreting, responding and reflecting. A third set of variables was created that averaged student performance across all 11 of the indicators within a specific simulation to identify an overall level of clinical judgment demonstrated, which facilitated comparisons between groups.

The student participants completed only one clinical judgment rubric at the conclusion of their participation in three simulation scenarios and associated debriefings. This clinical judgment self-assessment reflected an inherent summary of their performance. The student data could not be examined in relationship to a specific simulation scenario. This is a limitation of the study design but determined to be necessary in view of the limited facility time available, the numbers of students participating in the simulations and the additional time that would have been required for students to complete two additional rubrics.

The faculty raters completed three clinical judgment rubrics, one for each scenario. Data analysis from the faculty clinical judgment rating followed the same pattern; analysis by clinical judgment indicator, by dimension and by overall level of performance with a specific simulation. However, while faculty data were analyzed in relationship to a specific simulation scenario it was also aggregated across the three simulation scenarios. The term ‘summary’ when applied to the analyses represents data that were summarized across all three simulation scenarios. Table 3 presents a listing of terms and the associated method of data calculation.

Table 3.

Summary of Terms With Associated Method of Calculation

Term	Definition
Simulation Specific Terms	
Clinical judgment indicator score ▪ Focused observation ▪ Recognizing deviations from expected patters ▪ Information seeking	Represents the students’ score on one of the 11 indicators of clinical judgment in the rubric

Term	Definition
▪ Prioritizing data ▪ Making sense of data ▪ Calm, confident manner ▪ Clear communication ▪ Well-planned intervention/flexibility ▪ Being skillful ▪ Evaluation/Self-analysis ▪ Commitment to Improvement	
Clinical judgment dimension score ▪ Noticing ▪ Interpreting ▪ Responding ▪ Reflecting	One of four scores representing the average of the specific indicators associated with each of the dimensions.
Overall level of clinical judgment score ▪ Simulation #1 ▪ Simulation #2 ▪ Simulation #3	Single score representing the average score on all 11 indicators of clinical judgment; summarizes student level of performance in a specific simulation scenario
Aggregated Across Simulations	
Summary clinical judgment indicator score	Summarizes student performance on one of the 11 specific indicators of clinical judgment averaged across three simulation scenarios
Summary clinical judgment dimension score	Summarizes student performance on a particular dimension averaged across three simulation scenarios
Summary overall level of clinical judgment score	Summarizes the level of clinical judgment performance averaged across the three simulation scenarios

In all three hypotheses in this study, the outcome variable of interest was the student's level of clinical judgment performance. Table 4 presents a summary of the relationship between the research hypotheses, variables and method of data analysis. In the first hypothesis, the data were consistently compared by clinical judgment indicator, clinical judgment dimension and by overall level of performance demonstrated. The data analyzed were from the faculty raters. The two student groups were compared on each of the three simulation scenarios separately and then averaged across the three simulations for a summary clinical judgment score comparison between the two student groups. In the second hypothesis the source of the data used was faculty rating of clinical judgment performance. In this hypothesis, the focus was on the interaction between complexity and level of student rather than the previous framework. As the third hypothesis compared faculty assessment of clinical judgment to student self-assessment of performance, the source of the data was these two groups. The framework for analysis returned to clinical judgment indicator, dimension and overall level of performance.

Table 4.

The Relationship of the Hypotheses, Variables and Method of Data Analysis

Hypothesis	Variables	Number of Tests	Method of Analysis
1. Graduating seniors will demonstrate a higher level of clinical judgment than end-of-year juniors as rated by faculty and measured on the Lasater Clinical Judgment Rubric.	DV: Faculty LCJR scores by ▪ Summary clinical judgment indicator ▪ Summary clinical judgment dimension ▪ Summary overall clinical judgment score IV: student grouping ▪ End-of year juniors ▪ Graduating seniors	16	t-tests

Hypothesis	Variables	Number of Tests	Method of Analysis
2. As simulation complexity increases, student performance in clinical judgment, as measured on the Lasater Clinical Judgment Rubric, will decrease.	DV: Faculty LCJR score ▪ Summary overall clinical judgment score IV: two variables ▪ Student level ▪ Simulation complexity	3 1	t-tests Repeated measures ANOVA
3. There is a relationship between student self-evaluation of clinical judgment and faculty rating of that student's level of clinical judgment	DV: the LCJR scores by ▪ Summary clinical judgment indicator ▪ Summary clinical judgment dimension ▪ Summary overall clinical judgment score IV: grouping ▪ Faculty ▪ Students	16 2	Paired t-tests Repeated measures ANOVA

This data analysis plan required multiple applications of the t-test statistic. Multiple comparisons using this statistic are known to increase the chance of type I error (Stockburger, 2007). This risk of type I error increases when applied to items that are correlated and, in theory, all the dimensions in the clinical judgment rubric should correlate. The alpha level used to determine statistical significance was set at .05. Rather than simply reducing the alpha level in order to reduce the risk of type I error, the exact p values were reported. While the number of t-tests required in this study were large, all were necessary in order to achieve the goals of the work. An examination of the eleven specific indicators of clinical judgment was essential as this data provides the most helpful information to students. The clinical judgment dimension data was the source of information most helpful for faculty and in future clinical judgment research.

As an additional aid to the comparisons of the means between groups, effect sizes were calculated. According to the design of the rubric, the junior students were expected to perform at the developing level, scored as a ‘2’, and the senior students were expected to be at the accomplished level, scored as a ‘3’. The expected effect size differences between the two student groups are therefore expected to be large. This study represents the initial examination of the construct validity of the clinical judgment rubric so the use of exact p values and effect size facilitates a deeper examination of the relationships between the variables. In the next section, chapter four, the results from the application of this data analysis plan will be reported.

CHAPTER FOUR

RESULTS

The primary focus of this study was the examination of the construct validity of the Lasater Clinical Judgment Rubric in the performance evaluation of nursing students. Known-groups methodology was applied in this study to determine if the clinical judgment rubric was able to differentiate between two groups of nursing students known to have differing levels of ability. The setting for this study was a high-fidelity simulation theater. To facilitate clarity within this chapter, the presentation of the findings will be organized by research questions. However, since group equivalence on characteristics other than clinical judgment is central to the application of the known groups method, this chapter will begin with an examination of group characteristics.

Group Characteristics

The characteristics of the two groups of students, end-of-year juniors and graduating seniors, were examined with the goal of determining if the two groups were related on some characteristic other than the variable of interest, which was clinical judgment. The demographic variables of particular interest across these two groups included age, GPA on admission to nursing school, prior experience in caregiving, years of post high school education, gender, and previous academic degree. The comparisons between the two groups on the continuous variables were analyzed by t-test (Table 5). The comparisons between the two groups on the dichotomous variables were analyzed using chi-square (see Table 6.). There were no significant differences between the two groups of students on any of the potentially confounding variables.

Table 5.

Comparisons Between Juniors and Seniors on Age, GPA and Experience

	Juniors N=22		Seniors N=24			
	Mean SD)	Range	Mean (SD)	Range	t-stat	sig
Age	27.2(7.0)	20 - 57	27.5 (7.3)	21 - 51	-.210	.834
GPA on admission	3.53(0.23)	2.89 - 4.02	3.53 (0.23)	3.10 - 3.95	-.071	.943
Years of experience caregiving	1.64(1.99)	0 - 7	2.29(2.33)	0 - 11	-1.021	.313
Years of post high school education	2.23(2.20)	0 – 8	2.04(1.43)	0 - 4	.342	.734

Table 6.

Comparisons Between Juniors and Seniors on Gender and Previous Education

	Juniors N=22		**Seniors N=24**			
	No	Yes	No	Yes	X^2	sig
Male Gender	100% (22)	0% (0)	96% (23)	4% (1)	.937	.522
Prior 2 year degree	82% (18)	18% (4)	75% (18)	25% (6)	.314	.422
Prior 4 year degree	86% (19)	14% (3)	96% (23)	4% (1)	1.296	.271
Prior graduate degree	96%(21)	4% (1)	100% (24)	0% (0)	1.115	.478

Effect of testing order

Students selected their appointment times for participation in the high-fidelity simulation component of the study. As the simulations were a required class learning activity, 15 days of facility time were needed to accommodate all 72 students in both classes. The primary concern in testing order was to determine if the simulation scenario information was presented consistently to all students participating. Two potential sources of variation were identified. Within each day of testing, six students were scheduled, so it was possible that those scheduled in the morning were different from those scheduled in the afternoon; from either fatigue on the part of the simulation staff or fatigue on the part of the student participant. To examine this source of potential variations, student participants were split into two groups, morning and afternoon for comparison. A second potential source of variation was how consistently the simulation scenarios were implemented over the 15 days they were presented. To examine the data for consistency over the duration of implementation, students were grouped into scheduling quartiles to determine if simulation implementation changed over days of testing.

Regression analysis was conducted to determine if the two variables of time of day or quartile of testing had any impact on clinical judgment performance. These two variables, together, could only account for 4.3% of the overall variance in clinical judgment performance. Neither time of day of testing, nor quartile of testing was a statistically significant contributor to variance in level of clinical judgment performance (F (2,42) = .937, p = .40).

Study Findings by Research Question

Research Hypothesis #1: Graduating seniors will demonstrate a higher level of clinical judgment than end-of-year juniors as rated by faculty and measured on the Lasater Clinical Judgment Rubric.

Data from the four faculty raters were used to examine this hypothesis. Each student was rated by one rater, with the exception of the six student DVDs that were rated twice as part of the longitudinal inter-rater reliability assessment. The data were examined first through analysis of the eleven summary clinical judgment performance indicators, then by four summary clinical judgment dimensions and then by the three overall levels of clinical judgment scores from each simulation. Focusing the analysis on the data collapsed across student performance on all three of the simulation scenarios removed variation in student performance that was content specific. Effect sizes were reported in each of these analyses as well as statistical significance to facilitate the comparisons of the two groups of students.

The first analysis applied to this data compared the end-of-year junior to the graduating seniors using the summary clinical judgment indicator scores (see Table 7). For all eleven indicators in the Lasater Clinical Judgment Rubric, a statistically significant difference was found between the performances demonstrated by these two groups of students. On the Lasater Clinical Judgment Rubric, a score of 1 reflects a student at the beginning level, a score of 2 reflects the developing level, a score of e reflects accomplished performance and a score of 4 is exemplary. The faculty scores for the junior group tended to cluster at the borderline between the beginning and developing levels with clinical judgment scores ranging from 1.90 to 2.21. The faculty scores for the seniors placed them well into the developing level of clinical judgment, approaching the accomplished level with their scores ranging from 2.56 to 2.85.

Table 7.

Comparisons of Groups on Clinical Judgment Indicators as Rated by Faculty

Lasater Clinical Judgment Rubric - Indicator	Juniors Mean	Juniors SD	Seniors Mean	Seniors SD	*t*	*p*	Effect Size	z score
Noticing								
Focused Observation	1.98	.66	2.58	.76	-2.79	.008**	.86	81%
Recognizing Deviations from Expected Patterns	2.00	.69	2.57	.81	-2.51	.016*	.76	78%
Information Seeking	2.13	.79	2.60	.76	-2.05	.046*	.60	73%

Lasater Clinical Judgment Rubric - Indicator	Juniors Mean	Juniors SD	Seniors Mean	Seniors SD	*t*	*p*	Effect Size	z score
Interpreting								
Prioritizing Data	1.98	.71	2.67	.75	-3.12	.003**	.96	83%
Making Sense of Data	1.90	.66	2.58	.79	-3.09	.003**	.93	83%
Responding								
Calm, Confident Manner	2.32	.72	2.81	.75	-2.22	.031*	.66	75%
Clear Communication	2.06	.75	2.75	.75	-3.06	.004**	.92	82%
Well-Planned Interventions / Flexibility	1.98	.76	2.58	.91	-2.37	.022*	.71	76%
Being Skillful	2.17	.67	2.75	.66	-2.89	.006**	.87	81%
Reflecting								
Evaluation / Self-Analysis	2.21	.71	2.85	.67	-3.10	.003**	.93	82%
Commitment to Improvement	2.17	.70	2.81	.72	-2.97	.005**	.90	82%

* $p < .05$
** $p < .01$

Effect sizes were calculated for these data to provide another gauge of the magnitude of the differences between the two groups of students. Essentially, effect size reflects the number of standard deviations that separate the two groups. The guidelines used for evaluation of effect size were those developed by Cohen (1988) as follows:

- Effect size = 0.2 is small
- Effect size = 0.5 is medium
- Effect size = 0.8 is large

The effect size difference was in the moderate range on the four indicators of recognizing deviations from expected patterns and information seeking as well as responding in a calm, confident manner with a well-planned intervention. In the remaining seven indicators, the effect size differences between the two groups were large. A z-score probability interpretation was calculated from the effect size to identify the percentage of juniors that scored lower than the average score attained by the senior group. The smallest difference between the two groups was on the clinical judgment indicator of information seeking and on this indicator 73% of the juniors scored less than the average of the senior group. On the remaining clinical judgment indicators the percentage of juniors that scored less than the average of the senior group ranged from 75% to 83%.

The second analysis applied to the faculty data compared juniors to seniors on the dimensions of clinical judgment: noticing, interpreting, responding and reflecting. Summary data from all three simulations were used for this comparison. There were statistically significant differences between graduating seniors and end-of-year juniors on all four dimensions of the clinical judgment (see Table 8).

Table 8.

Comparison of Groups on Clinical Judgment Dimension as Rated by Faculty

Lasater Clinical Judgment Rubric - Dimension	Juniors		Seniors				Effect Size	z-score
	Mean	SD	Mean	SD	*t*	*p*		
Noticing	2.03	.68	2.58	.76	-2.54	.015*	.76	78%
Interpreting	1.94	.68	2.63	.76	-3.15	.003**	.94	83%
Responding	2.13	.69	2.72	.73	-2.77	.008**	.83	80%
Reflecting	2.19	.67	2.82	.69	-3.14	.003**	.93	82%

* $p < .05$
** $p < .01$

Effect sizes were also calculated for this data. Using Cohen’s (1988) guidelines these effect sizes were large, reflecting sizable differences between the two groups. However, the effect size calculations show that the student means were closer on the dimension of noticing and farthest apart on the dimensions of interpreting and reflecting. Z-score probability interpretation was also calculated for this data to summarize the percentage of juniors who scored less than the average of the senior group. Across the dimensions of clinical judgment the percentage of juniors who scored less than the average of the senior group ranged from 78 to 83 percent.

The final analysis applied to the faculty data compared juniors to seniors on their summary overall level of clinical judgment performance. The overall level of clinical judgment performance summarizes student level of performance in each of the three simulation scenarios. The data were aggregated across all three simulations to mitigate the influence of content specific knowledge. The t-test comparison found a statistically significant difference between the two groups of students on their summary overall level of clinical judgment (see Table 9). Faculty summary overall level of clinical judgment placed the junior students just at the entry of the developing level of clinical judgment with their summary score at 2.08 (*SD* = .67). Faculty placed the graduating seniors well into the developing level with a summary overall level score of 2.69 (*SD* = .72). The effect size difference between these two groups was large and the z-score probability interpretation states that 81% of the juniors scored less than the average of the senior group in level of clinical judgment performance across the three simulations.

Table 9.

Comparisons of Groups on Summary Overall Level of Clinical Judgment as Rated by Faculty

Lasater Clinical Judgment Rubric Score	Juniors		Seniors				Effect Size	z-score
	Mean	SD	Mean	SD	*t*	*p*		
Summary Overall Level of Clinical Judgment Simulations #1 - #3	2.08	.67	2.69	.72	-2.91	.006**	.87	81%

* $p < .05$
** $p < .01$

To summarize the analyses applied to hypothesis one, the data supports accepting the conclusion that graduating seniors demonstrate a higher level of clinical judgment than end-of-year juniors as rated by faulty and measured on the Lasater Clinical Judgment Rubric. The difference was present in all eleven of the clinical judgment indicators, in all four dimensions of the clinical judgment rubric and remained when performance was summarized across the content present in the three different situations. Faculty viewed end-of-year juniors as performing at the borderline between the beginning and the developing stages of clinical judgment. Faculty viewed graduating seniors as performing strongly in the developing stage, tending toward accomplished.

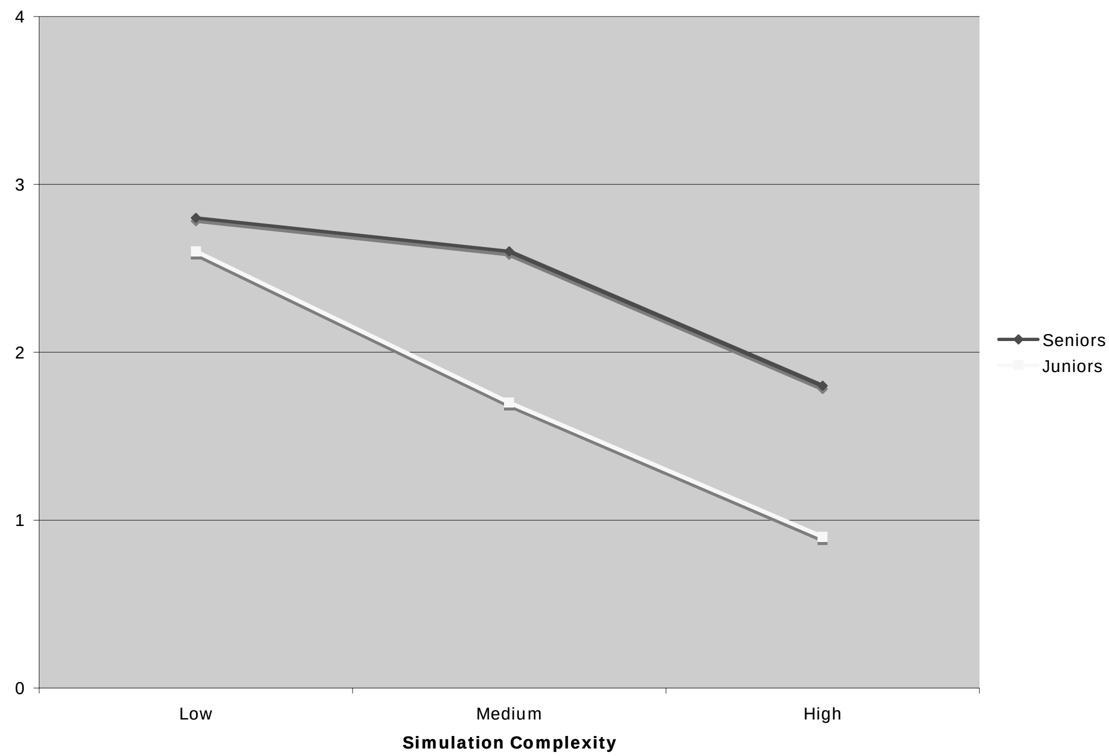
Study Question Two

Research Hypothesis #2. As simulation complexity increases, student performance in clinical judgment, as measured on the Lasater Clinical Judgment Rubric, will decrease.

An important aspect of this study was to explore the relationship between simulation complexity and the student's ability to demonstrate clinical judgment. One of the underlying assumptions of the research-based model of clinical judgment was that contextual complexity is necessary to differentiate novice clinical judgment from expert. For example, in a situation involving a stable patient and requiring only the assessment of vital signs, a beginning nurse could perform as well as an expert at this specific task. However, as complexity increases, if the patient's vital signs are not within normal limits, or those vital signs are not stable, the beginning nurse is not expected to be able to make the type of clinical judgments that an expert nurse would be able to make. Figure 3 presents the expected relationship among these variables.

Figure 3.

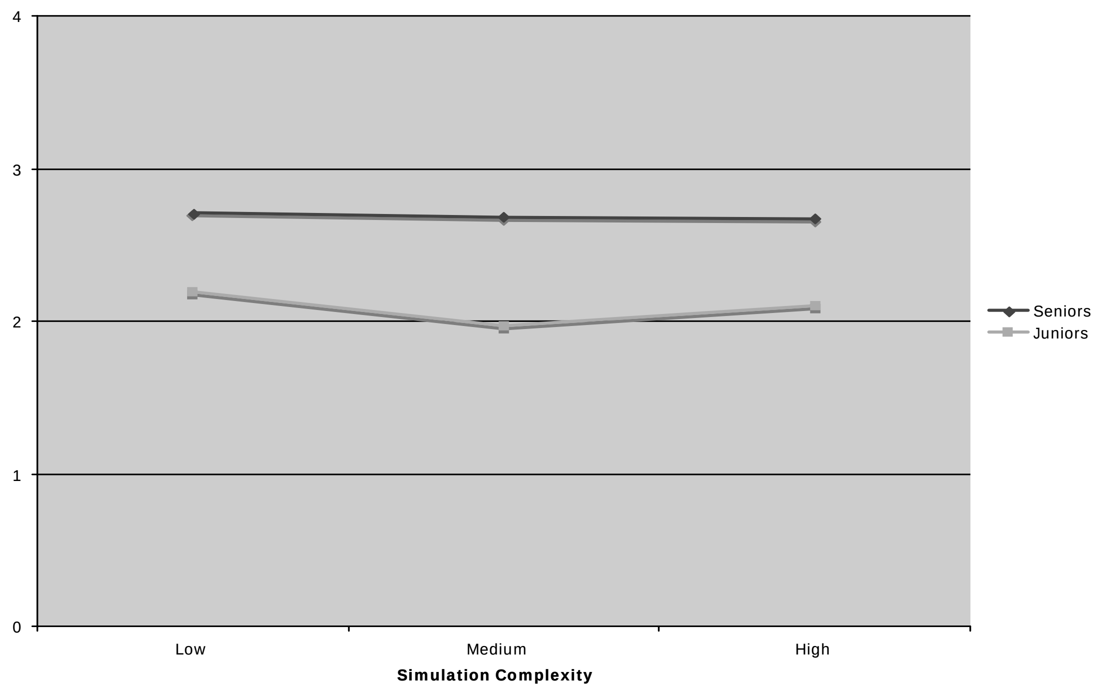
Expected Relationship Between Simulation Complexity, Student Level and Clinical Judgment



This study was designed to present students with three simulations of increasing complexity. To determine if an interaction was present between simulation complexity and student level on clinical judgment ability, a repeated measures analysis of variance was conducted using the data from the faculty raters' assessment of student clinical judgment. The three levels of complexity comprised the within subjects effects. The level of student grouping, as end-of-year junior or graduating senior, comprised the between subjects effect. The dependent variable of interest was the level of performance of clinical judgment as rated by faculty. The interaction (Figure 4) was not significant between student level and simulation complexity (F (2,86) = .634, p = .533). This indicates that faculty did not identify a statistically significant different pattern of clinical judgment performance between juniors and seniors as simulation complexity increased.

Figure 4.

Actual Relationship Between Simulation Complexity, Level of Student and Clinical Judgment



As expected, the data did demonstrate a significant difference in level of performance from juniors to seniors in all three simulations (see Table 10). However, a change in performance as complexity increased was not present. While both student groups did perform best in the low complexity simulation #1, there was little variation in clinical judgment performance rating in the subsequent two simulations.

Table 10.

Summary Comparisons of Overall Level of Clinical Judgment Performance

Lasater Clinical Judgment Rubric Score	Juniors		Seniors				Effect Size	z-score
	Mean	SD	Mean	SD	*t*	*p*		
Overall Level of Clinical Judgment Simulation #1	2.17	.71	2.71	.73	-2.57	.013*	.75	77%
Overall Level of Clinical Judgment Simulation #2	1.92	.75	2.68	.86	-3.16	.003**	.94	83%
Overall Level of Clinical Judgment Simulation #3	2.10	.76	2.67	.77	-2.47	.018*	.74	77%

* $p < .05$
** $p < .01$

To further examine the interaction of complexity and clinical judgment performance of students, the faculty raters were asked to give their opinion regarding the complexity of the three scenarios. In addition to evaluating simulation complexity along six dimensions, faculty was also asked to give a global complexity rating for each of the scenarios. To evaluate complexity, an 18 point visual line with equally spaced indicators of low, moderate and high complexity was developed (Appendix G). Scores that ranged from 1 to 6 reflected faculty perception of low complexity, scores from 7 to 12 indicated moderate complexity and scores from 13 through 18 were high complexity. Faculty were also asked to rate the global level of complexity of each simulation on this visual line.

The mean global complexity scores for the simulations were all within the moderate range (see Table 11). Faculty did not perceive the three simulations as presenting the full range of complexity. On an 18-point scale, faculty raters 1 and 4 found a five point difference and faculty raters 2 and 3 found a four point difference.

Table 11.

Global Complexity Rating as Rated by Faculty

	Simulation #1	Simulation #2	Simulation #3
Rater One	Low (5)	Moderate (8)	Moderate (10)
Rater Two	Moderate (10)	Moderate (12)	Moderate (14)
Rater Three	Low (4)	Moderate (7)	Moderate (8)
Rater Four	Moderate (10)	Moderate (13)	High (15)
Mean Value	Moderate (7.3)	Moderate (10.0)	Moderate (11.2)

To summarize the findings regarding research hypothesis two, there was no significant interaction between level of simulation complexity and level of student on ability in clinical judgment. However, it must be noted that faculty raters did not perceive the three simulations as presenting a full range of complexity. Possible explanations for this lack of interaction and alternatives will be explored in chapter five.

Study Question Three

Research Hypothesis #3: There is a relationship between student self-evaluation of clinical judgment and faculty rating of that student's level of clinical judgment.

The relationship between student self-evaluation of clinical judgment and faculty rating of that student's clinical judgment was explored in a series of correlational analyses. The summary data from the four faculty raters were used and compared to the single self-assessment completed by the students. A two-tailed Pearson r was the test statistic applied. The correlational data appears as a column rather than the typical diagonal because the faculty –to-student correlations were the focus of analysis rather than the faculty-to-faculty or student-to-student data, which were not reported. The data were initially examined by clinical judgment indicator, then by dimension and finally by overall level of clinical judgment performance.

The relationship between faculty rating of clinical judgment and student self-assessed level of performance was initially examined by clinical judgment indicator (see Table 12). The meaning of the correlations was interpreted against Cohen's (1988) guidelines:

- r = 0.10 – 29 is small
- r = 0.30 - .49 is medium
- r = 0.50 – 1.0 is large

Only three clinical judgment indicators demonstrated a significant correlation between faculty and student ratings. These indicators were focused observation, information seeking, and calm confident manner. The strength of these relationships was moderate.

Table 12.

Correlations Between Faculty and Student Ratings of Level of Clinical Judgment Performance

Clinical Judgment Indicator	*r*	sig.
Focused Observation	.32*	.03
Recognizing Deviations from Expected Patterns	.14	.35
Information Seeking	.32*	.03
Prioritizing Data	.17	.26
Making Sense of Data	.03	.85
Calm, Confident Manner	.31*	.04
Clear Communication	.25	.09
Well-Planned Intervention / Flexibility	.04	.81
Being Skillful	.13	.39
Evaluation / Self-Analysis	-.08	.61
Commitment to Improvement	-.04	.81

* $p < .05$
** $p < .01$

The next set of correlational analyses examined the relationship between faculty rating and student self-assessment of performance by clinical judgment dimension. Again, these correlations were presented in a column versus diagonal table to remove repetitive data that were not the focus of the question (see Table 13). The pattern present in clinical judgment indicators was clustered and this impacted the dimensions. The dimension of noticing demonstrated the only significant correlation. The students' self-assessment of their noticing skills correlated at a moderate level with the faculty rating of the student performance.

Table 13.

Correlations Between Faculty and Student Ratings of Clinical Judgment by Dimension

Clinical Judgment Dimension	*r value*	Sig.
Noticing	.31*	.04
Interpreting	.13	.40
Responding	.24	.11
Reflecting	-.05	.75

* $p < .05$
** $p < .01$

To summarize, the faculty assessment of the students' self-assessment of performance on clinical judgment correlated at a moderate level on the indicators of focused observation, information seeking, and calm confident manner. At the level of clinical judgment dimension, only noticing demonstrated a significant correlation between faculty rating and student self-assessment and the strength of the correlation was also moderate.

The relationship between faculty rating of student clinical judgment and the students' self-assessment of performance was further investigated through paired t-test analysis. The data compared were the students' clinical judgment indicator scores to the faculty summary clinical judgment indicator scores. Student groups were maintained for this analysis. There were statistically significant differences between faculty rating of clinical judgment and junior student self-assessment of performance in all eleven of the clinical judgment indicators (see Table 14). There were statistically significant differences between faculty rating of clinical judgment and senior student self-assessment of performance in ten of eleven clinical judgment indicators.

Table 14.

Faculty to Junior Student Comparisons of Clinical Judgment by Indicator

Lasater Clinical Judgment Rubric Indicator	Faculty		Juniors				Effect Size	z-score
	Mean	SD	Mean	SD	*t*	*p*		
Noticing								
Focused Observation	1.98	.66	2.88	.55	-5.35	.001**	-1.50	93%
Recognizing Deviations from Expected Patterns	2.00	.69	2.95	.67	-4.24	.001**	-1.40	92%
Information Seeking	2.12	.80	3.11	.50	-5.08	.001**	-1.52	94%
Interpreting								
Prioritizing Data	1.98	.71	2.98	.72	-4.54	.001**	-1.39	92%
Making Sense of Data	1.90	.66	2.83	.58	-3.88	.001**	-1.50	93%
Responding								
Calm, Confident Manner	2.32	.72	3.12	.48	-4.21	.001**	-1.33	91%
Clear Communication	2.06	.75	3.21	.64	-5.30	.001**	-1.64	95%
Well-Planned Interventions / Flexibility	1.98	.76	3.00	.72	-4.00	.001**	-1.38	92%
Being Skillful	2.17	.67	2.91	.37	-3.68	.001**	-1.42	92%
Reflecting								
Evaluation / Self-Analysis	2.21	.71	2.93	.51	-3.20	.004**	-1.18	88%
Commitment to Improvement	2.17	.70	3.33	.58	-4.97	.001**	-1.81	96%

* $p < .05$
** $p < .01$

As noted earlier, faculty tended to view the junior group of students as just past the beginning level of clinical judgment and barely entering into the developing stage. The student self-perception of their work placed them at the accomplished level of clinical judgment. In effect sizes calculation, the faculty score was entered first, then the student score resulting in the negative number. A pooled standard deviation was used. The large effect size results reveal the magnitude of the differences between the faculty assessment of clinical judgment and the junior student's self-perception of their clinical judgment ability. The z-score interpretation provides a framework for understanding the magnitude of difference between faculty rating and student self-assessment as this column identifies the percentage of junior students who rated themselves above the mean faculty rating for a specific clinical judgment indicator.

The faculty to senior students clinical judgment comparisons (see Table 15) continued to show the trend of student self-perception of ability as higher than that of the faculty. On ten of the eleven indicators, students' self-assessment was higher than faculty to a statistically significant degree. The indicator that did not show a significant difference between faculty and senior students was the indicator of evaluation/self-analysis in the dimension of reflecting. This trend toward improvement in self-analysis was echoed in the effect size results, which although large, were not as large as those seen in the junior group.

Table 15.

Faculty to Senior Student Comparisons of Clinical Judgment by Indicator

Lasater Clinical Judgment Rubric Indicator	Faculty		Seniors				Effect size	z-score
	Mean	SD	Mean	SD	*t*	*p*		
Noticing								
Focused Observation	2.58	.76	3.33	.56	-4.31	.001**	-1.14	87%
Recognizing Deviations from Expected Patterns	2.57	.81	3.31	.59	-4.04	.001**	-1.06	86%
Information Seeking	2.60	.76	3.54	.51	-6.05	.001**	-1.47	93%
Interpreting								
Prioritizing Data	2.67	.75	3.29	.82	-3.00	.006**	-0.78	78%

Lasater Clinical Judgment Rubric Indicator	Faculty		Seniors				Effect size	z-score
	Mean	SD	Mean	SD	*t*	*p*		
Making Sense of Data	2.58	.79	3.08	.48	-3.16	.004**	-0.78	78%
Responding								
Calm, Confident Manner	2.81	.75	3.56	.54	-5.22	.001**	-1.15	87%
Clear Communication	2.75	.75	3.42	.58	-4.44	.001**	-1.02	85%
Well-Planned Interventions / Flexibility	2.58	.91	3.29	.67	-3.24	.004**	-0.90	82%
Being Skillful	2.75	.66	3.19	.44	-3.27	.003**	-0.80	79%
Reflecting								
Evaluation / Self-Analysis	2.85	.67	3.21	.66	-1.83	.081	-0.54	71%
Commitment to Improvement	2.81	.72	3.58	.58	-4.26	.001**	-1.18	88%

* $p < .05$
** $p < .01$

A comparison of the effect sizes and the z-scores associated was more informative than the tests of statistical significance in describing these two groups. The faculty to junior student effect sizes ranged from -1.33 to -1.81. The faculty to senior student effect sizes ranged from -0.54 to -1.47. The effect size comparison presents the senior group as moving closer in their self-assessment to the position of the faculty raters. The calculation of a z-score interpretation changed the effect size into a probability and shows the percentage of students that rated themselves higher than the faculty mean. The z-scores for the senior group, while still large show a smaller percentage of senior students rated themselves more highly than the faculty mean on an indicator.

A repeated measures analysis of variance was applied to this data to determine if the pattern of closer alignment between faculty rating and student self-assessment as students moved from junior level to senior level was significant. The within subjects factors were the faculty to junior student and faculty to senior student. The between subjects factor was the student grouping, junior or senior. The analysis of variance was conducted on the individual clinical judgment indicators (see Table 16) as well as the summary overall clinical judgment level of clinical judgment.

Table 16

Analysis of Variance Between Faculty and Student Rating of Clinical Judgment by Indicator

Clinical Judgment Indicator	*F* statistic $_{(1,43)}$	*p* value
Noticing		
Focused Observation	.36	.549
Recognizing Deviations in Expected Patterns	.53	.471
Information Seeking	.05	.826
Interpreting		
Setting Priorities	1.47	.231
Making Sense of Data	2.34	.134
Responding		
Calm, Confident Manner	.081	.778
Clear Communication	3.50	.068
Well-Planned Intervention / Flexibility	.85	.362
Being Skillful	1.57	.218
Reflecting		
Evaluation / Self-Analysis	1.47	.233
Commitment to Improvement	1.70	.200

* $p < .05$
** $p < .01$

This analysis of variance demonstrated that student grouping did not significantly impact the accuracy of self-assessment in clinical judgment; the seniors were no more accurate than the juniors in self-evaluation on any of the specific clinical judgment indicators. When the data used was from the overall level of clinical judgment to summarize the total performance, the ANOVA results remained insignificant (*F* (1,43) =

1.54, p = .235). However, given the limitations of sample size in this study, this test may have lack lacked sufficient statistical power to evaluate this interaction.

This paired t-test analysis examined the relationship between specific pairs of faculty and students and demonstrated a significant difference was present between faculty perception of clinical judgment and student self-assessment. From the initial analysis of t-test results it appeared that the differences between faculty and junior students and faculty and senior students became narrower over time. However, repeated measures ANOVA testing did not support the difference between student groupings as significant. There are many possible explanations for the difference in application of the clinical judgment rubric between faculty and students. These possibilities will be addressed in chapter five.

CHAPTER FIVE

DISCUSSION AND CONCLUSIONS

The purpose of this study was examination of the construct validity of the Lasater Clinical Judgment Rubric in the context of high-fidelity simulation. Construct validity refers to the extent to which evidence and theory support the interpretation of the scores resulting from a proposed use of the instrument (*Standards*, 1999). The theoretical relationships present in the research-based model of clinical judgment provided two testable assumptions for the examination of construct validity (1) that nurses with greater practical experience and greater levels of nursing specific knowledge would demonstrate higher levels of clinical judgment; (2) that the level of clinical judgment ability demonstrated would be related to the level of situational complexity present. Empirically testing these two assumptions formed the core of this validity study. While construct validity is currently viewed as a unitary concept, five different types of validity evidence are acknowledged (*Standards*, 1999). This study in particular examined three of the five types of evidence. The study design, using the known groups approach examined the evidence based on relations to other variables. In support of the use of the known groups approach, a second type of evidence, the evidence based on content was examined through a review of literature. The purpose of the review of literature was to explore the extent to which the evidence supports the view that clinical judgment is a developmental ability facilitated by exposure to nursing. The literature was also used to examine the evidence related to the content of the rubric – that is, showing the relationship of the rubric to the concept domain of clinical judgment, to its developmental indicators, thus supporting the proposed interpretation of the scores. The third type of validity evidence examined in this study was evidence based on response processes. As the intended use of the clinical judgment rubric was performance assessment, examination of the extent to which the raters were consistent in applying the rubric was important.

The second hypothesis of this study examined the relationship between clinical judgment performance ability and situational complexity. The relationship between complexity and clinical judgment is one of the underlying testable assumptions of the research-based model of clinical judgment. Specifically this study predicted that there would be an interaction among situational complexity, level of experience and clinical judgment ability. Current research addressing the relationship between task complexity and performance on measures of clinical judgment is limited. This study provided an empirical examination of these relationships.

In addition to the focus on construct validity, this study also served a secondary area of research interest, which was the exploration of the role of clinical judgment assessment in a learner-centered curriculum. This aim was explored through an examination of the relationship between student self-assessment and faculty assessment. While it is important for faculty to accurately assess the level of student ability, it is equally important for students to develop accurate self-evaluation. As clinical judgment is a developmental process the importance of appropriate self-assessment extends beyond the educational setting.

This chapter addresses the implications and limitations of the study findings. The construct validity evidence in three areas – content, relationship to other variables, and response process – will be examined first. Next, the second aspect of construct validation, the evidence relating to the level of situational complexity, level of nursing ability and performance of clinical judgment is presented. Following this, a discussion of the findings related to self-assessment of clinical judgment will be presented. From the exploration of the implications and limitations of this current study, areas for further research will be identified.

Construct Validity: The Evidence Based on Relations to Other Variables

Construct validation is an ongoing process that emerges from the development of a chain of evidence that logically links the foundational theoretical constructs with the interpretation of the assessment scores in a manner that supports the relevance of the interpretation. An examination of the evidence based on relations to other variables was a logical base from which to begin such a construct validity argument. The review of literature was the initial step in this process as it provided a means for examining the relevance of clinical judgment as a content domain appropriate for the evaluation of differences between levels of students. The literature review supported the position that clinical judgment is developmental and does change in the course of nursing education. This study compared end-of-year juniors in a baccalaureate nursing school and graduating seniors on their level of clinical judgment performance ability. The support for validity from evidence based on relations to other variables is present to the extent the study found the expected differences in clinical judgment ability between these two groups of students. The evidentiary support for clinical judgment as an appropriate domain for the evaluation of differences between students will be further examined against Tanner's (2000, 2006) research-based model of clinical judgment.

One of the conclusions that Tanner articulated from the summary of research that lead to the development of her model of clinical judgment was that "clinical judgments are more influenced by what the nurse brings to the situation than the objective data about the situation at hand" (2006, p205). This conclusion supports the position that clinical judgment is developmental. Nurses that can bring greater theoretical and practical knowledge to the encounter with a patient will demonstrate a higher level of ability. What gets noticed in any situation is dependent on both what the nurse brings to the situation, as well as the situation itself. As the context for this study was the setting of high-fidelity simulation, the objective data about the situation at hand was equalized. Across the study, all participants received identical preparatory information and worked through the same three patient scenarios. Thus, the theoretical assumption regarding the influence of nursing knowledge and experienced was tested by comparing the performances of two groups of students, end-of-year juniors and graduating seniors. It was expected that graduating seniors with their advanced ability gained from their coursework and clinical practica would demonstrate a higher level of clinical judgment.

The performance assessment data obtained in this study through the application of the LCJR by raters' blind to the educational level of the students found the difference between the two groups predicted by Tanner's research-based model. The students participated in three different simulation scenarios and these ratings were averaged to remove effect of specific content knowledge and provide an accurate reflection of student knowledge. The mean clinical judgment score faculty raters assigned to the end-of-year juniors was 2.08 (*SD* = .67), placing this group at the borderline just barely past beginning level 1. The mean clinical judgment score faculty raters assigned to the graduating seniors was 2.69 (*SD* = .72). This rating places the senior students on the LCJR solidly in the developing level 2 stage and working toward the accomplished level

3. The LCJR as an assessment tool was able to identify the expected presence of two groups. The differences identified between the two groups of students were statistically significant in all of the eleven indicators of clinical judgment.

To summarize, construct validity evidence based on relations to other variables examines ‘the degree to which these relationships are consistent with the construct underlying the proposed test interpretation” (*Standards*, p13). One of the constructs underlying the Tanner research-based model of clinical judgment was the assumption that greater nursing knowledge and practical experience would influence the level of clinical ability demonstrated. Findings from this study support the construct validity of the Lasater Clinical Judgment Rubric as the expected differences predicted by Tanner’s research-based model of clinical judgment were discernable from application of the rubric. The senior nursing students brought a greater level of knowledge and experience to the clinical situation than the junior nursing students and their advanced abilities were apparent in the clinical judgments demonstrated when measured on the Lasater Clinical Judgment Rubric.

Construct Validity: The Evidence Based on Content

According to the *Standards* (1999), evidence based on content includes “logical or empirical analyses of the adequacy with which the test content represents the content domain and of the relevance of the content domain to the proposed interpretation of test scores” (p. 11). The content present in the Lasater Clinical Judgment Rubric was based on the conceptual framework of the Tanner research-based model of clinical judgment. The dimensions of the model provided the framework for the subsequent development of the eleven clinical judgment indicators. Lasater used “a cycle of theory-driven description-observation-revision-review" (2007, p498) to develop each of the behavioral descriptors present in the rubric. To examine the construct validity evidence that is based on content, the rubric descriptors will be examined using the structure of the underlying model and the proposed interpretations will be compared to evidence from the literature.

Discussion regarding the dimension of noticing

The first specific theoretical link to examine in the chain of construct validity evidence is a comparison of the literature that pertains to noticing in relationship to the findings on the LCJR. The literature supports the positive influence of nursing education on the student's ability to notice. Beginning with the study by Verhonic, Nichols, Glor and McCarthy (1968), and continuing with the studies by Davis (1972; 1974), education was shown to have a positive impact on nurse's ability to make relevant observations and take action. Later studies on this topic (Botti & Reeves, 2003; Thiele et al., 1991; Tschikota, 1993) found education to have a positive influence on the student's ability to differentiate relevant from irrelevant cues. The summary of the literature supports the ability to differentiate relevant from irrelevant cues as a hallmark feature of developing clinical judgment in nursing.

On the LCJR, the dimension of noticing explicitly examines the issue of cue relevancy through the indicator of focused observation, but also addresses the influence of the nurse's knowledge base on what is noticed as well as the ability to extract information in the setting. In the indicator of focused observation, the LCJR identifies a beginning level 1 student as confused and disorganized by the situation, while the developing level 2 student is able to focus on the most obvious data but remains overwhelmed by the array of data, and the accomplished level three student is able to observe and monitor most of the useful information, missing only the most subtle signs. The indicator of recognizing deviations from expected patterns addresses the extent of the nurse's knowledge base and the ability to access that knowledge base in the process of noticing relevant cues about the patient. The indicator of information seeking addresses how effectively a nurse seeks out information and addresses cue relevancy by determining if the nurse fails to collect information at the beginning level 1, or pursues unrelated information at developing level 2, or occasionally does not pursue an important lead at accomplished level 3. These three indicators on the LCJR fully embody the concept of salience and are descriptive of a nurse's pattern of growth in this process.

In this study, the LCJR was able to differentiate between the two groups of students in their abilities to notice. On the dimension as a whole, the junior students had a mean clinical judgment rating by faculty of 2.03 (*SD* = .68), or just at the developing level. The senior students had a mean clinical judgment rating of 2.58 (*SD* = .76). The results show the greatest difference between juniors and seniors related to their abilities in focused observation, then their abilities in recognizing deviations from expected patterns. The indicator demonstrating the least difference between the groups was information seeking. When comparing this to the other dimensions of interpreting, responding and reflecting, the difference between the two groups was the smallest in the area of noticing, but the z-score interpretation of the effect size still indicates that 78% of the juniors scored lower than the average score of the senior group.

The senior group mean in the dimension of noticing was rated at 2.58 (*SD* = .76) and is still within the developing level 2. However, the score differences between the two groups sets the key level differentiating this group of students at the choice between developing level 2 and the accomplished level 3. The senior group was significantly different from the junior group and closer to level 3. This finding supports the position that seniors were more able to observe a variety of useful information, missing the subtle signs while the juniors tended to focus on the most obvious data and were more likely to be overwhelmed by the array of data. The seniors were more likely to recognize patterns and deviations and continually assess the patient while the juniors were more likely to recognize only the most obvious pattern, and miss important information. The seniors were more likely to actively seek out information and only occasionally fail to pursue important leads, while the juniors were limited in seeking out additional information, hampered by not knowing what information to seek and more often pursued unrelated information.

The theoretical relationships present in the research-based model of clinical judgment supports the pattern of differences found between these two groups. In Tanner's model, three factors frame the nurses' expectations and set up their initial grasp of what is noticed within a situation. These factors are the relationship between the nurse and the patient, the situational context and, most importantly, the background of the nurse including their theoretical and practical knowledge. The least difference found between the juniors and the senior students was on the clinical indictor of information seeking (ES = .60). This was not unexpected as it is one of the earliest skills taught and practiced in a nursing curriculum. Finding the greatest difference within this dimension between the two groups on the clinical indicators of focused observation (ES = .86) followed by recognizing deviations from expected patterns (ES = .76) was also not unexpected. Nursing capability on these two clinical judgment indicators requires both the acquisition of domain-specific knowledge and the ability to apply it in clinical practice. The senior students were able to perform better because they had acquired more background knowledge and had more experience in the clinical context.

Discussion regarding the dimensions of interpreting and responding

The next theoretical link to examine in the chain of construct validity evidence is a comparison of the literature that pertains to decision making in relationship to the findings on the LCJR. One of the strengths of the research-based clinical judgment model is the inclusion of multiple alternative types of reasoning patterns. The choice of reasoning pattern varies with the background of the nurse and the immediate situation. The results of the decision made are visible in the response pattern made by the nurse and the outcomes seen in the patient. Thus with clinical judgment performance assessment, the topic is broader than a focus on any one type of decision making process or on whether or not the 'correct' decision was made. In clinical judgment performance assessment the focus is on the dynamic of how the data are interpreted and prioritized and how the nurse uses the information to respond to patient problems.

From the literature there are two hallmark characteristics that differentiate the inexperienced from the experienced nurse in the areas of interpreting and responding. In the absence of an acquired base of practical knowledge, inexperienced nurses predominantly use an analytic pattern of reasoning (Benner, 2004). An analytic pattern allows the inexperienced nurse to match factual cues from a clinical situation to their classroom-acquired knowledge base. As the student's knowledge base expands, so does their ability to interpret this knowledge base in relationship to their patient (Tschikota, 1993) and apply their knowledge base in the development of patient specific responses

(Corcoran, 1986). The second hallmark characteristic that continues to differentiate level of clinical judgment relating to interpreting and responding is cue relevance. The difficulty an inexperienced nurse has with cue relevancy impacts the ability to interpret data and set priorities. If all cues are perceived as equally relevant it is difficult for the student to establish patterns in the data and set priorities (Thiele et al, 1991). The amounts of clinical practice experience, as well as the extent of the student's theoretical knowledge base, were found to impact the ability to differentiate cue relevance (Botti and Reeves, 2003).

The characteristic of cue relevance as a means of differentiating clinical judgment ability in interpreting and responding in the literature is present in the descriptors on the LCJR. There is a definite application of relevance to the clinical judgment indicator of prioritizing data in the dimension of interpreting. Clinical judgment capability in prioritizing data is defined in terms of the nurses' ability to attend to identify and attend to relevant data. This sense of relevance is also present in dimension of responding in the indicator labeled well-planned intervention/flexibility. Within this indicator, nursing capability is defined by the ability to identify an appropriate solution as well as the ongoing ability to maintain a pertinent plan.

The other hallmark characteristic identified in the literature, the extent and ability to use a nursing knowledge base is also visible in both dimensions. The indicator of making sense of data, in the dimension of interpreting, clearly addresses the extent of the nurses' knowledge base as a key factor. However, knowledge base is also present as an underlying theme in the dimension of responding. For an intervention to be a well-planned intervention, it should reflect a current knowledge base. The remaining two clinical judgment indicators in the dimension of responding, calm, confident manner and nursing skillfulness, were developed from Lasater's (2005) observations of students and not from the research-based model of clinical judgment.

In this study, the LCJR was able to differentiate between the end-of-year juniors and the graduating seniors on their ability to interpret as part of their clinical judgment assessment. The faculty assessed the junior group ability at interpreting data at 1.94 (SD = .68) placing them just under the borderline of the developing level 2. Faculty assessed the senior group ability at interpreting data at 2.63 (SD = .76) placing them well up into the developing level 2. The difference is statistically significant and the effect size of 0.94 demonstrates the distance between the means of these two groups as almost an entire standard deviation apart. The z-score interpretation of this effect size indicates that 83% of the junior group was rated lower than the average of the senior group.

Of all the dimensional analyses, the differences between the two groups of students were largest in the area of interpretation. On both the indicators in this dimension, prioritizing data and making sense of data, the seniors were significantly more advanced than the juniors. This matches what is known from the literature (Botti and Reeves, 2003; Thiele, 1991; Tschikota, 1993) regarding the expansion of a student's knowledge base through exposure to the nursing curriculum and the ability to use that knowledge to differentiate relevant cues.

The LCJR, in this study, was also able to differentiate between the two groups on the dimension of responding. The faculty assessed the junior group clinical judgment ability in responding at 2.13 (*SD* = .69) and the senior group at 2.72 (*SD* = .72), placing both groups within the developing level 2 but situating the senior group as significantly higher in ability. Of the four indicators in this dimension, two emerged from Lasater's (2005) observational work, those of calm, confident manner and skill in nursing therapeutics. The remaining two indicators, clear communication and well-planned intervention / flexibility have strong links to the theoretical model and link directly back to the dimension of interpreting so require specific discussion.

Interpreting must begin as an internal process and only becomes visible either through the process of responding or, if the nurse chooses, to reflect-in-action aloud. It is through the responses by the nurse of communication and interventions that provides insight to the observer concerning the student's interpretation of data. The clinical judgment indicator that demonstrated the greatest difference between the groups was clear communication (ES = .92). Senior students were more accomplished in the ability to effectively establish rapport and provide clear explanations to patients. This supports the position that a deeper understanding of the patient's health care alteration, acquired through education, facilitated the students' ability to interpret data and communicate their understanding of pertinent issues. There was also a significant difference between seniors and juniors on the indicators of well-planned intervention / flexibility. On this indicator, both the accomplished level 3 nurse and the developing level 2 nurse were able to conduct ongoing monitoring of the patient's progress, but it was the accomplished nurse who maintained an expectation of needing to change treatments. This expectation emerges from the accomplished nurses' deeper understanding of alternative interpretations.

To summarize, senior students demonstrated a more advanced ability on both the interpreting and responding dimensions of clinical judgment performance. The seniors possess a larger knowledge base, both theoretical and practical, and this positively influenced their abilities to interpret relevant data, and set priorities. The junior group was more likely to attend to less relevant data and have difficulty making sense of patient data patterns. The senior group was also more likely to respond with clear communication, a broader range of relevant interventions and perform psychomotor skills more proficiently.

Discussion regarding the dimension of reflecting

The final theoretical link to examine in the chain of construct validity evidence is a comparison of the literature that pertains to reflection in relationship to the findings on the LCJR. In the literature, reflection holds a central position in the development of understanding. From Dewey (1910) through Schon (1983) to Mezirow (1991) to King and Kitchener (1994) these theorists all believe reflection is essential to the development of ability. To narrow the focus to nursing, the literature does support that reflection improves thinking strategies (Fonteyn & Cahill, 1998; Smith, 1998). The literature further supports that although barriers to reflection are present in the educational system, these can be overcome and reflection can be accurately assessed from student written work (Boenink et al. 2004; Pee et al. 2000, 2002; Wong et al., 1995). Finally, the literature supports that reflective ability does change over time as it both facilitates and demonstrates the underlying change in domain-specific ability of the nursing student (Kuiper, 2002; Murphy, 2004).

The LCJR is well designed to capture the potential differences in reflective ability. The influence of Benner's body of work regarding the development of nursing expertise (1984, 1996, 2004) and the influence of King and Kitchener's work relating to the development of reflective judgment (1994) are clearly visible. The dimension of reflection has two indicators, evaluation/self-analysis and commitment to improvement. The descriptions at the beginning level present someone who is rarely reflective, whose justifications for action are minimal as the individual functions from rule-driven understanding of the nursing domain. As reflective ability develops, the LCJR levels describe the nurse as someone who gradually requires less external prompting to engage in evaluation and self-analysis. Depth of analysis regarding decisions made and numbers of alternatives considered also expand as reflective ability improves. Within the indicator of commitment to improvement, depth of analysis is also a key component in determining the individual's level of reflective ability. The LCJR describes the individual moving from the position of being non-reflective at the beginning level 1 to an individual who is metacognitive at the exemplary level 4.

In this study, the LCJR was able to differentiate between junior and senior students in relationship to their ability on the dimension of reflecting. The faculty rated the junior student ability on reflection at 2.19 (*SD* = .67) placing them at the developing level 2. The faculty raters assessed the senior group ability in reflection at 2.82 (*SD* = .69), placing them also in the developing level 2, but at a much higher level of performance. While this difference in ability is statistically significant, it is also important to note the large effect size of 0.93 demonstrates the distance between the means of the two groups are almost one standard deviation apart. In this analysis, the senior group required less prompting to identify key decisions and discuss alternatives while the junior group tended to focus on the obvious and had difficulty developing alternative choices. While both groups demonstrated a commitment to improvement, the senior group was more specific in their identification of personal strengths and weaknesses, while the junior group again, tended to state the obvious and require external prompting to self-reflect.

The difference found between these two groups on the reflecting dimension is consistent with the literature pertaining to the development reflective ability. As students advance in the knowledge base and in their level of clinical experience in nursing, their ability to reflect on their actions improves. The group of senior students in this study demonstrated better ability in evaluation of their nursing actions, in their assessment of personal performance, and in the ability to development of a plan for improvement.

Construct Validity: the Evidence Based on Response Process

A type of construct validity evidence important in this study relates to the response processes of the faculty raters. According to the *Standards* (1999) one aspect of response process evidence concerns the "extent to which the processes of observers or judges are consistent with the intended interpretation of scores" (p. 13). In this study the response process of interest is the extent to which the raters were consistent with the intended use of the LCJR in their scoring of students. The potential threat to construct validity of concern is the presence of construct-irrelevant variance. The *Standards* (1999) define construct-irrelevant variance as systematic error affecting the assessment data that originates in variables unrelated to the construct of interest. The level of agreement between the faculty raters changed over the course of this study, indicating construct-irrelevant variance influenced the study findings. The threat of construct-irrelevant variance as revealed by inter-rater reliability levels is the focus of this section.

The inter-rater level of agreement was initially high. At the completion of the training process, all raters attained a greater than 90% level of agreement in the application of the LCJR to the low complexity simulation scenario. By the end of the study, after the completion of thirteen rounds of student evaluations, rater levels of agreement had diminished significantly. This raises a serious concern. The conventional view holds reliability as a necessary, but not sufficient, condition for validity (Nunnally & Bernstein, 1994). However, concluding the analysis of the evidence as being un-interpretable secondary to low reliability at this point would be inappropriate. To understand the weight of the evidence relating to responses processes, a deeper exploration of the issues is needed.

Inter-rater reliability was calculated in this study as level of agreement. For each clinical judgment indicator on the LCJR, the scores of the two raters were compared to determine if they agreed or disagreed on the level of ability displayed by that student. The final score reflected the percentage of agreement between the raters on the eleven indicators. This process was repeated for each of the three simulations and then averaged. The final inter-rater level of agreement was a summary of how well the two raters had agreed on the eleven rubric indicators in each of the three simulations.

Calculating inter-rater reliability using level of agreement in this situation emerged as problematic. A fundamental statistical assumption of reliability analyses is that the items are reasonably independent (Downing, 2003; Downey & Haladyna, 2004). Variables nested in sets, such as the indicators nested in the LCJR, violate the assumption of independence. In this situation, the unit of reliability analysis is the case. Obviously, using only two raters and providing only three cases was insufficient data and this increased the likelihood of finding low levels of agreement. When the standard of agreement between raters was expanded by one point, a process that adjusted the reliability analysis in response to the nested nature of the clinical judgment indicators, the inter-rater reliability remained satisfactory over the course of the study.

The limited reliability findings present in this study are likely related to the methods used in the rater training. Faculty raters were provided two types of training. The first type was rater error training, which was provided for the purpose of making raters more aware of potential errors associated with rubric use. The errors presented in the training focused on those most common as identified by Boston (2002), such as leniency, halo effect, central tendency, and skimming, among others. The second type of training was performance dimension training. This type of instruction was provided to train raters in the dimensions of clinical judgment performance so they would recognition the appropriate level of each dimension during their observations of performance.

When multiple faculty assess the same student and arrive at different conclusions, there are two possible types of rater bias present (Williams, Klamen & McGaghie, 2003). One option is that different raters focus on, or differentially weight, different aspects of the performance. This is commonly known as trait bias and it combines easily with the halo effect. For example, faculty may view the student's proficiency with physical assessment as most important and the student ability (or lack of ability) influences the rating in other areas. Another option is that different raters have different expectations concerning acceptable levels of performance. This is most commonly seen as the leniency and stringency effect. Rater error training has been reviewed as moderately effective in decreasing halo error and in some cases increasing rater accuracy (Woehr &

Huffcutt, 1994). However, the deficiency with the rater error training provided in this study was the time lag between the instruction regarding the potential errors and the start time of the faculty actually rating performance assessments for the study. The time lag varied among the raters from 4 to 6 months. This time lag is a study limitation as it mitigates any beneficial effect that may have resulted from the original training.

The second type of rater training provided in this study was performance dimension training. This process started with the raters watching and discussing each of the three simulation scenarios in order to learn the dimensions and levels of the Lasater Clinical Judgment Rubric. After this initial day of interactive practice, when the inter-rater reliability results were not sufficient, the focus of the training shifted. In order to replace a lost rater and accommodate the disparate work schedules and geographic locations of the raters, the training method altered from group discussion to independent modules. In addition, instead of ongoing practice with all three of the scenarios, practice was directed toward one scenario only. Logically, using the performance dimension model of rater training, the expectation was that faculty would acquire an understanding of the dimension of clinical judgment exhibited in one situation and transfer this understanding to two other simulation scenarios. However with the move to independent training modules, the component of active group discussion was lost.

In performance dimension training, the focus is on the identification of particular dimensions of the construct, in this case clinical judgment. The drawback to performance dimension training is that it does not increase accuracy (Woehr & Huffcutt, 1994). The training developed faculty that could identify the performance dimensions to clinical judgment but without the group interaction, there was no group agreement regarding specific levels of performance. There is support in the literature for the use of independent modules as a type of gold standard training (Williams, Klamen & McGaghie, 2003) and this proved effective with this group of faculty initially. However, the lack of ongoing feedback and without reinforcement for the levels of clinical judgment, faculty raters drifted apart.

There are multiple recommendations present in the literature to address low inter-rater reliability. The first recommendation would be to institute frame of reference training, which teaching raters to identify performance standards as well as performance dimensionality (Woehr & Huttcutt, 1994). In this study, the four faculty raters were provided only with same the simulation scenario preparatory information that was given to the students, which included a brief history and physical and the admitting physician orders. While the simulation case progression was developed from the model of clinical judgment, identifying what cues were planned for the students to notice, followed by optional responses that varied with how they interpreted and responded to the cues, it is important to note this information was not shared with the faculty raters. There was no a priori discussion of the goals and objectives of the simulation learning activity. Frame of reference research (Williams, Klamen & McGaghie, 2003) suggests this is the single most effect method for improving accuracy of observation and rating. Sharing the simulation case progression, which identifies situational options for the assessment of the indicators of clinical judgment, with the faculty raters, is a logical place to begin frame of reference training. A second component of frame of reference training would be to provide faculty raters with specific examples of students performing at the different levels of clinical judgment. These types of examples would serve to anchor the faculty rater's expectations for performance. This study did not provide faculty raters with this type of anchoring information and must be considered a limitation.

A second recommendation would be to provide the faculty raters with feedback on the consistency of their rating during the process (Williams, Klamen & McGaghie, 2003). This would serve to identify when faculty need to refresh their frame of reference and diminish the tendency of a leniency-stringency error. The third recommendation would be to continue with rater error training but to minimize the time between training session and actual rating work.

To summarize, the construct validity evidence based on response process is equivocal. When calculated most conservatively, the inter-rater reliability levels between the faculty raters are unacceptably low. The inconsistency of the faculty raters diminishes the evidence of a valid difference between the student groups. However, consideration of the nested nature of the clinical judgment indicators is reasonable and expansion of the level of agreement by one point is a credible statistical compensation. With this adjustment, the inter-rater levels of agreement are supportive of the evidence based on response process. Overall, the lessons learned from the process of training faculty raters need to be brought forward and the recommendations for improvement integrated into future research.

Implications Regarding Simulation Complexity and Clinical Judgment

This study began an investigation of how complexity is defined in the high fidelity simulation environment. The goal was to examine student performance across a set of increasingly complex situations. The purpose was to examine the interaction between complexity and level of student on clinical judgment performance. The theoretically testable assumption was the expectation that as situational complexity increased, clinical judgment performance would decrease. However, in this study such an interaction was not found. Simulation complexity did not have an impact on the level of clinical judgment demonstrated in either group of students.

The lack of an interaction between simulation complexity and clinical judgment ability was a surprising finding. Experts are those who are characterized as people who not only possess a broad foundational knowledge base but also can extract the salient information present in a complex situation fluently and rapidly (Benner, 2004; Bransford, 2000). Complexity should impact the demonstration of ability through sheer situational volume of potential cues alone. Before the potential interaction between complexity and student level on clinical judgment is rejected, a thorough examination of the nature of complexity as defined in this study is required.

There are four possible explanations for the lack of interaction in this study between complexity and clinical judgment. (1) an error in the measurement of complexity; (2) an error in the scenario implementation of the chosen level of complexity; (3) the possibility of a practice effect resulting in improved student performance from the first simulation through the third; and (4) insufficient power within the study design.

Measurement of simulation complexity was initiated through the development of a simulation complexity rubric. Three levels of complexity were defined, high, medium, and low. Six dimensions to the construct of complexity were identified; learner pre-knowledge of the situation, task/setting complexity, physiological complexity, psycho/social/spiritual complexity, information source complexity, and potential risk/danger to SimPatient complexity. An investigation of the perceptions of the faculty raters concerning their opinion of the complexity present in the three simulation scenarios revealed they did not see distinctly different levels of complexity present. Faculty viewed complexity as clustered in either the low to moderate range or in the moderate to high range. No faculty thought the simulations represented the spectrum of complexity from low to moderate to high. Faculty viewed these three simulations as having a restriction in range. Thus, it is apparent the three levels of complexity in the present rubric need to be refined to reflect further differentiation of level. The low complexity level needs to be defined more simply and the high complexity level needs to become significantly more complex.

Errors in the implementation of scenario complexity emerged from inadequate description of each of the dimensions of complexity. While each of the three simulations was designed to reflect a particular level of complexity, faculty raters perceived a wide variation in the complexity present within specific dimensions. For example, at the scenario design stage, the dimension of physiological complexity seemed to be a relatively straightforward dimension varying with the acuity of the patient's presentation and number of co-morbidities. However, during simulation implementation, within one simulation scenario, faculty rated this aspect of complexity from low to the moderate-high range. Clearly, the dimensions of complexity require additional work to further explicate and refine the definitions. However, while faculty made suggestions to refine the definitions of the various dimensions of complexity, no faculty identified a component of complexity that was not already present in the six dimensions.

A third option for the lack of interaction is the possibility of a practice effect. Students moved directly from the first simulation and its debriefing right through to the second and then the third. The advantage of simulation education is the engagement of the student in the process of active learning. It is possible that the student acquired knowledge during each of the simulations, even under the pressure of a performance evaluation, and used this improved understanding in their subsequent work.

A fourth possible explanation for not finding an interaction between complexity and level of student is an insufficiently powerful study design. Only four faculty raters provided the data and their opinions concerning the complexity present varied markedly. The data are an insufficient basis for any conclusions. Accurate identification of simulation complexity is important. If complexity can be effectively differentiated in the setting of simulation then scaffolded learning experiences can be provided to students. Appropriately scaffolded learning experiences support retention of learning. If the learning activity is too easy, the student does not acquire new understandings. If the learning activity is too difficult, then the student not only does not acquire new understandings but also ends up confused. To maximize the effectiveness of the student’s time spent in simulation, the learning activities need to be appropriately challenging. Identification of simulation complexity is the first step. Further research is therefore essential concerning both simulation complexity and its relationship to clinical judgment.

Comparison of Faculty to Student Clinical Judgment Evaluations

Faculty to student clinical judgment evaluations were compared by three methods. The first method used correlational analysis, the second, and a paired t-test and the third, repeated measures analysis of variance. From a performance assessment standpoint, high, positive correlations between faculty rating and student self-assessment would have been optimal. A high, positive correlation would represent symmetry between faculty expectations of clinical judgment performance and the students' understanding of those expectations. However, the correlational analyses conducted in this study revealed faculty rating of clinical judgment correlated at a moderate level with student self-assessment only on the indicators of focused observation, information seeking and calm, confident manner. At the dimension level faculty and students scores only correlated in the dimension of noticing and did not correlate in the dimensions of interpreting, responding or reflecting.

The second method of comparing faculty to student clinical judgment evaluation used the paired t-test. In these analyses, faculty and students differed in their perceptions of the level of clinical judgment demonstrated with the student self-evaluations being significantly higher. Faculty assigned the senior group a mean clinical judgment score of 2.69 (SD = .72) while the seniors self-assessed their clinical judgment to be 3.35 (SD = .44). The junior group echoed this pattern with the faculty average clinical judgment score at 2.08 (SD = .67) and the junior self-assessed mean at 3.02 (SD = .36).

There are several possible causes for these differing perceptions. First, neither the juniors nor the senior student had received any formal training in the use of the clinical judgment rubric. While both groups of students had previously been exposed to the concept of clinical judgment and the model describing the process had been explained and applied in their prior work in simulation, neither group of students had specific practice using the rubric. Students had not been provided with an opportunity to discuss the language used in the rubric and clarify meaning. Instructions to students were simply to read the rubric and circle the components they felt were the best descriptors of their performance. One possible cause of the difference between student and faculty could simply be a reflection of student unfamiliarity with the rubric language.

Another possible explanation for the differences between the perceptions of faculty and students could be related to measurement issues within the rubric. The dimension that correlated best between faculty and students was noticing. This dimension has the more easily externalized indicators of clinical judgment that in essence measure what the nurse can say and do. The dimensions of interpreting and reflecting, that did not correlate between faculty and student, can be more difficult to evaluate as they are more representative of internal thought processes.

The examination of the differences between faculty and student assessment of clinical judgment requires deep investigation. Theoretically, a nurse cannot interpret and respond to data that has not been first noticed. For example, under the dimension of noticing and within focused observation the rubric identifies the accomplished nurse as someone who 'regularly observes/monitors a variety of data, including both subjective and objective; most useful information is noticed, may miss subtle signs'. Students can lack an awareness of the subtle signs and therefore do not recognize important aspects have been missed. An accomplished nurse watching will see the deficiency, as they possess a broader understanding. In information seeking, an accomplished nurse 'occasionally does not pursue important leads'. Inexperienced nurses who lack either cognitive or practical knowledge have difficulty recognizing important leads have been missed, where a more experienced nurse will notice this type of relevant cue promptly.

Unless the information that is not noticed triggers a subsequent alteration in patient status, the beginning nurse will continue forward under the impression that everything important has been noticed and all pertinent information has been obtained.

The conceptual model of learning in simulation presented in chapter two of this paper viewed feedback as one of he most important factors influencing learning. The literature supports this view. Feedback is essential to the development of reflection, to giving students high quality information about their performance in order to close the gap between present and future performance goals (Juwah et al., 2004). However, for feedback to fulfill these functions the learner must be actively engaged and the criteria of what constitutes a good performance must be clearly understood by the learners (Fink, 2003). Lasater (2005) found that students want clear and honest feedback about the quality of their performance in order to make the most effective improvements. The differences between faculty assessed level of clinical judgment and student self-assessed clinical judgment found in this study suggests that students do not, at present, have a clear understanding of what constitutes expert clinical judgment.

Summary and Conclusions

The process of construct validation first requires the theoretical relationships between the concepts be linked. This study started with two such testable assumptions: 1) Nurses with greater practical experience and greater levels of nursing specific knowledge will demonstrate higher levels of clinical judgment 2) Level of clinical judgment ability demonstrated will be related to the level of situational complexity present. The second step in construct validation is to examine the empirical relationships between measures of the construct. The Lasater Clinical Judgment Rubric provided the empirical data. As an observational tool, this rubric provided a means for the evaluation of student clinical judgment in response to the specific situation represented in the simulation scenario. The third step of construct validity is to interpret the empirical evidence as it relates back to the theoretical concepts.

The empirical evidence provided by the LCJR did not support the second assumption that level of clinical judgment ability demonstrated would be related to the level of situational complexity. This does not indicate that the assumption is invalid or the LCJR was ineffective in discerning a difference. At this point in the research process, the only firm conclusion that can be made from the data is that complexity is difficult to describe and measure. It is particularly difficulty to measure high-fidelity complexity that is an attempt to reflect realistic challenges that nurses encounter. However, the control and replication that is possible in simulation support this as the location for further research toward defining complexity.

The empirical evidence provided by the LCJR does support the assumption that nurses with greater practical experience and greater levels of nursing specific knowledge demonstrate higher levels of clinical judgment. The empirical data provided by the Lasater Clinical Judgment Rubric found senior nursing students were able to demonstrate significantly higher levels of clinical judgment then junior nursing students. While the threat of construct-irrelevant variance manifested in the modest inter-rater reliability findings does limit the interpretation of the empirical support, strong and significant differences were present in these two groups of nursing students. The empirical evidence supported the theoretical link between the concepts.

Recommendations for Future Research

The first recommendation for future research is to identify a method to improve inter-rater reliability. Specifically, new research needs to focus on the addition of frame of reference training for the raters. This frame of reference training needs to include thorough discussion among the rating faculty the simulation case progression. The case progression should identify the planned cues for students to notice, and possible options for interpretation and response. In addition to the active component, case progression should also include the essential debriefing questions that prompt student reflection on key aspects. Using frame of reference training with faculty should also include providing the resource of performance demonstrations that establish each of the different levels of student ability. The effectiveness of using performance anchors to improve inter-rater reliability is an important area for future research.

A second for future research is to further explore the construct of complexity in the high-fidelity simulation setting. This is an important priority because it is foundational to the construct validity of the rubric. It is also important component for any future research synthesis of data. A valid complexity rubric would facilitate comparisons across sites using other topics for simulation but at equivalent levels of complexity so it would be helpful in comparing students at benchmark points. In a similar vein, an understanding of simulation complexity would be helpful in comparing outcomes from nursing curriculums. With a valid clinical outcome measure and a valid complexity instrument, curricular points could be identified and comparisons made. Finally, understanding complexity will be an important part of any attempt to determine how much simulation is present in a curriculum and when attempting to establish a research link between performance in simulation and performance in clinical practice.

A third recommendation for future research is to explore ways in which we can develop within students the ability to be metacognitive regarding their clinical judgment level. Metacognition by students does not seem to emerge accurately simply as a result of direct reflection-on-action. This study suggests that development of effective metacognitive ability requires practice specific to the development of that skill. The development of a context specific nursing roles to frame the students' practice of clinical judgment would be one method of enhancing clarity of expectations. Other avenues of research might be to examine the influence of case-based classroom work prior to simulation practice to determine if this facilitates student understanding of the expected level of clinical judgment performance. Conversely, the method of allowing the student to enter the simulation setting for their initial practice of clinical judgment and then providing either written or video demonstrations of the expected level of clinical judgment ability is an option for future research.

The use of high-fidelity simulation as a method of nursing education is expanding rapidly. A reliable and valid outcome measure to determine the effectiveness of this educational method and more importantly to evaluate the performance capabilities of nurses is essential. This study demonstrates the Lasater Clinical Judgment Rubric to be a robust nursing outcome measure. The data in this study supports the construct validity of the rubric with the caution that further research needs to be undertaken to improve consistency in application as well as deepening the understanding of situational complexity. This study demonstrates the Lasater Clinical Judgment Rubric is an outcome measure that is worth such a research investment.

References

Adams, B. L. (1999). Nursing education for critical thinking: An integrative review. *Journal of Nursing Education, 38*(3), 111-119.

Adams, M. H., Whitlow, J. F., Stover, L. M., & Johnson, K. W. (1996). Critical thinking as an educational outcome: An evaluation of current tools of measurement. *Nurse Educator, 21*(3), 23-32.

Aiken, L.H., Clarke, S.P., Sloane, D.M., Sochalski, J., Silber, J.H. (2002). Hospital nurse staffing and patient mortality, nurse burnout and job dissatisfaction. *Journal of the American Medical Association, 288*(16), 1987-1993

American Association of Colleges of Nursing. (1998). *The Essentials of baccalaureate education for professional nursing practice.* Washington, DC: Author.

American Educational Research Association. American Psychological Association. National Council on Measurement in Education. (1999*). Standards for Educational and Psychological Testing.* Washington, D.C.: Author

American Psychological Association Task Group of the Board of Educational Affairs. (1997, November). Learner-centered psychological principles: A framework for school reform and redesign. Washington, DC: Author.

Banta, T. W. (2001). Assessing competence in higher education. In C. A. Palomba & T. W. Banta (Eds.), *Assessing student competence in accredited disciplines: Pioneering approaches to assessment in higher education* (pp. 1-12). Sterling, VA: Stylus.

Barr, R. B., & Tagg, J. (1995). From teaching to learning - A new paradigm for undergraduate education. *Change*, 27(6), 12-26.

Bearnson, C. S., & Wiker, K. M. (2005). Human patient simulators: A new face in baccalaureate nursing education at Brigham Young University. *Journal of Nursing Education, 44*(9), 421-425.

Beckie, T. M., Lowry, L. W., & Barnett, S. (2001). Assessing critical thinking in baccalaureate nursing students: A longitudinal study. *Holistic Nursing Practice, 15*(3), 18-26.

Benner, P. (1984). *From novice to expert: excellence and power in clinical nursing practice*. Menlo Park, CA: Addison-Wesley.

Benner, P. (2004). Using the Dreyfus Model of skill acquisition to describe and interpret skills acquisition and clinical judgment in nursing practice and education. *Bulletin of Science,* 24(3), 188-199.

Benner, P., Tanner, C. A., & Chesla, C. A. (1996). *Expertise in nursing practice: Caring, clinical judgment and ethics*. New York: Springer.

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

Benner, P., Tanner, C.A., Chesla, C.A. (1997). The social fabric of nursing knowledge. *American Journal of Nursing, 97*(7), 16BBB-16DDD.

Bjork, I. T., & Kirkevold, M. (1999). Issues in nurses' practical skill development in the clinical setting. *Journal of Nursing Care Quality*, *14*(1), 72-84.

Boenink, A. D., Oderwald, A. K., De Jonge, P., Van Tilburg, W., & Smal, J. A. (2004). Assessing student reflection in medical practice. The development of an observer-rated instrument: reliability, validity and initial experiences. *Medical Education, 38*(4), 368-377.

Boland, D. L., & Laidig, J. (2001). Assessment of student learning in the discipline of nursing. In C. A. Palomba & T. W. Banta (Eds.). *Assessing student competence in accredited disciplines: Pioneering approaches to assessment in higher education* (pp. 71-94). Sterling, VA: Stylus.

Botti, M., & Reeve, R. (2003). Role of knowledge and ability in student nurses' clinical decision making. *Nursing and Health Sciences, 5*, 39-49.

Boston, C. (2002). *Understanding scoring rubrics: A guide for teachers*. ERIC Clearinghouse on Assessment and Evaluation. College Park, MD.

Boud, D., Keogh, R., & Walker, D. (1985). *Reflection: Turning Experience into Learning*. New York: Nichols.

Boulet, J. R., Murray, D., Kras, J., Woodhouse, J., McAllister, J., & Ziv, A. (2003). Reliability and validity of a simulation-based acute care skills assessment for medical students and residents. *Anesthesiology, 99*(6), 1270-1280.

Bowles, K. (2000). The relationship of critical-thinking skills and the clinical-judgment skills of baccalaureate nursing students. *Journal of Nursing Education, 39*(8), 373-376.

Boyd, E. M., & Fales, A. W. (1983). Reflective learning: Key to learning from experience. *Journal of Humanistic Psychology, 23*(2), 99-117.

Brannon, L. A., & Carson, K. L. (2003). Nursing expertise and information structure influence medical decision making. *Applied Nursing Research, 16*(4), 287-290.

Bransford, J.D., Brown, A., & Cocking, R.R. (Eds.). (2000). *How people learn: Brain, mind, experience and school* (Expanded ed.). Washington, D.C: National Academy Press.

Brooks, K. L., & Shepard, J. M. (1990). The relationship between clinical decision-making skills in nursing and general critical thinking abilities of senior nursing students in four types of nursing programs. *Journal of Nursing Education, 29*(9), 391-399.

Brown, H. N., & Sorrell, J. M. (1993). Use of clinical journals to enhance critical thinking. *Nurse Educator, 18*(5), 16-19.

Brown, D. M. (2003). Learner-centered conditions that ensure students' success in learning. *Education*, 124(1), 99-107.

Brunt, B. A. (2005). Critical thinking in nursing: An integrated review. *The Journal of Continuing Education in Nursing, 36*(2), 60-67.

Bucknall, T., & Thomas, S. (1997). Nurses' reflections on problems associated with decision-making in critical care settings. *Journal of Advanced Nursing, 25*, 299-237.

Bruner, J. (1986). *Actual Minds, Possible Worlds*. Cambridge, MA: Harvard University Press.

Byrne, A. J., & Jones, J. G. (1997). Responses to simulated anaesthetic emergencies by anaesthetists with different durations of clinical experience. *British Journal of Anaesthesia, 78*(5), 553-556.

Campbell, D.T. & Fiske, D.W. (1959). Convergent and discriminant validity in the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Carmines, E.G. & Zeller, R.A. (1979). *Reliability and Validity Assessment*. Series: Quantitative applications in the social sciences. London: Sage.

Chaiklin.S. (2003). The zone of proximal development in Vygotsky's Analysis of Learning and Instruction. In A. Kozulin, B. Gindis, V. S. Ageyev & S. M. Miller (Eds.), *Vygotsky's Educational Theory in Cultural Context*. Cambridge, MA: University Press.

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.

Cohen, J. (1988). *Statistical Power Analysis for Behavioral Sciences*. Hillsdale, New Jersey: Lawrence Erlbaum.

Colucciello, M.L. (1997). Critical thinking skills and dispositions of baccalaureate nursing students – A conceptual model for evaluation. *Journal of Professional Nursing, 13*(4), 236-245.

Corcoran, S. A. (1986). Planning by expert and novice nurses in cases of varying complexity. *Research in Nursing & Health, 9*, 155-162.

Cronbach, L.J. & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Davis, B. G. (1972). Clinical expertise as a function of educational preparation. *Nursing Research, 21*(5), 530-534.

Davis, B. G. (1974). Effect of levels of nursing education on patient care: A replication. *Nursing Research, 23*(2), 150-155.

Devitt, J. H., Kurrek, M. M., Cohen, M. M., Fish, K., Fish, P., Noel, A. G., et al. (1998). Testing internal consistency and construct validity during evaluation of performance in a patient simulator. *Anesthesia & Analgesia, 86*(6), 1160-1164.

Devitt, J. H., Kurrek, M. M., Cohen, M. M., & Cleave-Hogg, D. (2001). The validity of performance assessments using simulation. *Anesthesiology, 85*(1), 36-42.

Dewey, J. (1910). *How we think*. Mineola, NY: Dover.

Dewey, J. (1916). Democracy and education: An introduction to the philosophy of education. New York: Simon & Schuster.

Downey, S.M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education, 37*(4), 830-837.

Downey, S.M., & Haladyna, T.M. (2004). Validity threats: Overcoming interference with proposed interpretations of assessment data. *Medical Education, 38,* 327-333.

Ebright, P. R., Patterson, E.S., Chalko, B.A., & Render, M.L. (2003). Understanding the complexity of registered nurse work in acute care settings. *Journal of Nursing Administration, 33*(12), 630-638.

Ebright, P.R., Urden, L., Patterson, E., & Chalko, B. (2004). Themes surrounding novice nurse near-miss and adverse-event situations. *Journal of Nursing Administration, 34*(11), 531-538.

Ellermann, C.R., Kataoka-Yahiro, M.R., Wong, L.C. (2006). Logic models used to enhance critical thinking. *Journal of Nursing Education, 45*(6), 220-227.

Ericsson, K. A. (2004). Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Academic Medicine, 79*(1), S70-S81.

Facione, P.A. (1990). Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction. Research findings and recommendations. Newark, NJ: American Psychological Association.

Facione, P.A., Sanchez, C.A., Facione, N.C., & Gainen, J. (1995). The disposition toward critical thinking. *The Journal of General Education, 44*(1, 1-25.

Facione, N.C., & Facione, P.A. (1996). Externalizing the critical thinking in knowledge development and clinical judgment. *Nursing Outlook, 44,* 129-136.

Farrand, L. L., Holzemer, W. L., & Schleutermann, J. A. (1982). A study of construct validity: Simulations as a measure of nurse practitioners' problem-solving skills. *Nursing Research,* 31(1), 37-42.

Feingold, C. E., Calaluce, M., & Kallen, M. A. (2004). Computerized patient model and simulated clinical experiences: Evaluation with baccalaureate nursing students. *Journal of Nursing Education, 43*(4), 156-163.

Fink, L.D. (2003). Creating significant learning experiences: An integrated approach to designing college courses. San Francisco: Jossey-Bass.

Fonteyn, M. E., & Cahill, M. (1998). The use of clinical logs to improve nursing students' metacognition: A pilot study. *Journal of Advanced Nursing, 28*(1), 149-154.

Ford, J. S., & Profetto-McGrath, J. (1994). A model for critical thinking within the context of curriculum as praxis. *Journal of Nursing Education, 33*(8), 341-344.

Frederickson, K., & Mayer, G. G. (1977). Problem solving skills: What effect does education have? *American Journal of Nursing, 77*(7), 1167-1169.

Frye, B., Alfred, N., & Campbell, M. (1999). Use of the Watson-Glaser Critical Thinking Appraisal with BSN students. *Nursing & Health Care Perspectives, 20*(5), 253-255.

Gaba, D. M., Howard, S. K., Flanagan, B., Smith, B. E., Fish, K. J., & Botney, R. (1998). Assessment of clinical performance during simulated crises using both technical and behavioral ratings. *Anesthesiology, 89*(1), 8-18.

Gaba, D. M. (2004). A brief history of mannequin-based simulation & application. In W. F. Dunn (Ed.), *Simulation in critical care and beyond* (pp. 7-14). Des Plaines, IL: Society of Critical Care Medicine.

Goodwin, L.D. (1997). Changing conceptions of measurement validity. *Journal of Nursing*

Education, 36(3), 102-107.

Goodwin, L.D. (2002). Changing conceptions of measurement validity: An update on the new Standards. *Journal of Nursing Education 41*(3), 100-106.

Gordon, J. M. (2000). Congruency in defining critical thinking by nurse educators and non-nurse scholars. *Journal of Nursing Education, 39*(8), 340-351.

Gordon, M., Murphy, C.P., Candee, D., & Hiltunen, E. (1994). Clinical judgment: An integrated model. *Advances in Nursing Science, 16*(4), 55-70.

Henneman, E. A., & Cunningham, H. (2005). Using clinical simulation to teach patient safety in an acute/critical care nursing course. *Nurse Educator, 30*(4), 172-177.

Henrichs, B., Rule, A., Grady, M., & Ellis, W. (2002). Nurse Anesthesia students' perceptions of the anesthesia patient simulator: A qualitative study. *AANA Journal, 70*(3), 219-225.

Henry, S. B., LeBreck, D. B., & Holzemer, W. L. (1989). The effect of verbalization of cognitive processes on clinical decision making. *Research in Nursing & Health*, 12, 187-193.

Hertel, J. P., & Millis, B. J. (2002). *Using simulations to promote learning in higher education*. Sterling, VA: Stylus.

Hicks, F. D. (2001). Critical thinking: Toward a nursing science perspective. *Nursing Science Quarterly, 14*(1), 14-21.

Holzemer, W. L. (1986). The structure of problem-solving in simulations. *Nursing Research,* 35(4), 231-236.

Holzemer, W. L., Resnik, B., & Slichter, M. (1986). Criterion-related validity of a clinical simulation. *Journal of Nursing Education*, 25(7), 286-290.

Holzemer, W. L., & McLaughlin, F. E. (1988). Concurrent validity of clinical simulations. *Western Journal of Nursing Research*, 10(1), 73-83.

Institute of Medicine (2003). *Priority Areas for National Action: Transforming Health Care Quality*. Washington, D.C.: Author

Issenberg, S.B., McGaghie, W.C., Petrusa, E.R., Gordon, D. & Scalese, R.J. (2005). Features and uses of high-fidelity medical simulations that lead to effective learning: A BEME systematic review. *Medical Teacher, 27*(1), 10-28.

Itano, J. K. (1989). A comparison of the clinical judgment process in experienced registered nurses and student nurses. *Journal of Nursing Education, 28*(3), 120-126.

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.

Jeffries, P. R., & Rizzolo, M. A. (2004). *Designing and implementing models for the innovative use of simulation to teach nursing care of ill adults and children: A national, multi-site, multi-method study. Report of phase III*. New York: National League for Nursing.

Jeffries, P.R., & Rizzolo, M.A. (2006). *Designing and implementing models for the innovative use of simulation to teach nursing care of ill adults and children: A national multi-site, multi-method study. Summary report.* New York: National League for Nursing.

Jenks, J.M. (1993). The pattern of personal knowing in nurse decision making. *Journal of Nursing Education, 32*, 399-405.

Jenkins, H.M. (1985a). Improving clinical decision making in nursing. *Journal of Nursing Education, 24*(6), 242-243.

Jenkins, H. M. (1985b). A research tool for measuring perceptions of clinical decision making. *Journal of Professional Nursing*, 221-229.

Jenkins, H.M. (2001). Clinical decision making in nursing scale. In C.F. Waltz & L.S. Jenkins, Eds.), *Measurement of Nursing Outcomes* (2nd ed., pp 33-40). New York: Springer.

Jenny, J.J. & Logan, J. (1992). Knowing the patient: One aspect of clinical knowledge. *Image: The Journal of Nursing Scholarship*, *24*, 254-258.

Jones, S. A., & Brown, L. N. (1991). Critical thinking: Impact on nursing education. *Journal of Advanced Nursing*, *16*, 529-533.

Juwah, C., McFarland-Dick, D., Matthew, B., Nicol, D., Ross, D., & Smith, B. (2004). Enhancing student learning through effective formative feedback. Retrieved June 16, 2006 from The Higher Education Academy website: http://www.heacademy.ac.uk/assets/York/doc.

Kataoka-Yahiro, M., & Saylor, C. (1994). A critical thinking model for nursing judgment. *Journal of Nursing Education, 33*(8), 351-356.

Kautz, D.D., Kuiper, R., Pesut, D.J., Knight-Brown, P., & Daneker, D. (2005). Promoting clinical reasoning in undergraduate nursing students: Application and evaluation of the outcome present state test (OPT) model of clinical reasoning. *International Journal of Nursing Education Scholarship, 2*(1), 1-19.

Kennison, M. M., & Misselwitz, S. (2002). Evaluating reflective writing for appropriateness, fairness and consistency. *Nursing Education Perspectives*, *23*(5), 238-242.

King, P. M., & Kitchener, K. S. (1994). *Developing reflective judgment: Understanding and promoting intellectual growth and critical thinking in adolescents and adults*. San Francisco, CA: Jossey-Bass.

Kintgen-Andrews, J. (1991). Critical thinking and nursing education: Perplexities and insights. *Journal of Nursing Education, 30*(4), 152-157.

Kuiper, R. (2002). Enhancing metacognition through the reflective use of self-regulated learning strategies. *Journal of Continuing Education in Nursing, 33*(2), 78-87.

Kuiper, R. A., & Pesut, D. J. (2004). Promoting cognitive and metacognitive reflective reasoning skills in nursing practice: self-regulated learning theory. *Journal of Advanced Nursing, 45*(4), 381-391.

Larew, C., Lessans, S., Spunt, D., Foster, D., & Covington, B.G. (2007). Innovations in clinical simulation: Application of Benner's theory in an interactive patient care simulation. Nursing Education Perspectives, 27(1), 16-21.

Lasater, K. (2005). The impact of high fidelity simulation on the development of clinical judgment in nursing students: An exploratory study. (Doctoral Dissertation, Portland State University), (UMI No. AAI3169414), 227.

Lauri, S., & Salantera, S. (1998). Decision-making models in different fields of nursing. *Research in Nursing & Health, 21*, 443-452.

LeBreck, D.B. (1989). Clinical judgment: A comparison of theoretical perspectives. In W.L. Holzemer (Ed.), *Review of Research in Nursing Education* (pp33-55). New York: NLN.

Lederman, L.C. (1992). Debriefing: Toward a systematic assessment of theory and practice. *Simulation & Gaming, 23*(2), 145-160.

Leenders, M. (1998). Case research: The case writing process. Toronto: University of Ontario Press.

Lewis, M. L. (1997). Decision-making task complexity: Model development and initial testing. *Journal of Nursing Education, 36*(3), 114-120.

Long, R. E. (2005). Using simulation to teach resuscitation: An important patient safety tool. *Critical Care Nursing Clinics of North America, 17*(1), 1-8.

McCarthy, M.C. (2003a). Situated clinical reasoning: Distinguishing acute confusion from dementia in hospitalized older adults. *Research in Nursing and Health, 26*(2), 90-101.

McCarthy, M.C. (2003b). Detecting acute confusion in older adults: Comparing clinical reasoning of nurses working in acute, long-term, and community health care environments. *Research in Nursing and Health, 26*(3), 203-212.

McCarthy, P., Schuster, P., Zehr, P., & McDougal, D. (1999). Evaluation of critical thinking in a baccalaureate nursing program. *Journal of Nursing Education*, *38*(3), 142-144.

Messick, S. (1988). The one and future uses of validity: Assessing the meaning and consequences of validity. In H. Wainer & H.I. Braun (Eds), *Test validity* (pp33-45). Hillsdale, NJ: Lawrence Erlbaum.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Mezirow, J. (1991). *Transformative dimensions of adult learning*. San Francisco: Jossey-Bass Publishers.

Mezirow, J. (1994). Understanding transformation theory. *Adult Education Quarterly*, *44*(4), 222-232

Mezirow, J. (2000). *Learning as transformation: Critical perspectives on a theory in progress*. San Francisco, CA: Jossey-Bass.

Morgan, P. J., Cleave-Hogg, D., McIlroy, J., & Devitt, J. H. (2002). Simulation technology: A comparison of experiential and visual learning for undergraduate medical students. *Anesthesiology, 96*(1), 10-16.

Morse, J.M. (1992). Negotiating commitment and involvement in the nurse patient relationship. *Journal of Advanced Nursing, 16,* 455-468.

Mort, T. C., & Donahue, S. P. (2004). Debriefing: The basics. In W. F. Dunn (Ed.), *Simulation in critical care and beyond* (pp. 76-83). Des Plaines, IL: Society of Critical Care Medicine.

Murray, D., Boulet, J., Ziv, A., Woodhouse, J., Kras, J., & McAllister, J. (2002). An acute care skills evaluation for graduating medical students: A pilot study using clinical simulation. *Medical Education, 36*(9), 833-841.

Murray, D. J., Boulet, J. R., Kras, J. F., Woodhouse, J. A., Cox, T., & McAllister, J. D. (2004). Acute care skills in anesthesia practice: A simulation-based resident performance assessment. *Anesthesiology*, *101*(5), 1084-1095.

Murphy, J. I. (2004). Using focused reflection and articulation to promote clinical reasoning: An evidenced-based strategy. *Nursing Education Perspectives, 25*(5), 226-231.

National League for Nursing (1992). Criteria for evaluation of baccalaureate & higher degree programs in nursing (5th ed.). (NLN Publication No. 15-2474). New York: Author.

National League for Nursing Accrediting Commission (2006). *Accreditation Manual with Interpretive Guidelines by Program Type.* Retrieved August 1, 2007 from www.nlnac.org.

Nehring, W.M., Ellis, W.E., & Lashley, F.R. (2001). Human patient simulators in nursing education: An overview. *Simulation and Gaming, 32*(2), 194-204.

Nehring, W. M., & Lashley, F. R. (2004). Current use and opinions regarding human patient simulators in nursing education. *Nursing Education Perspectives, 25*(5), 244-248.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd edition ed.). New York: McGraw-Hill.

Oregon State Board of Nursing. (February, 2001). Nurse practice act: Division 21. Salem, OR: Author.

O'Sullivan, P.S., Blevins-Stephens, W.L., Smith, F.M. & Vaughan-Wrobel, B. (1997). Addressing the National League for Nursing critical-thinking outcome. *Nurse Educator, 22*(1), 23-29.

Pardue, S. F. (1987). Decision-making skills and critical thinking ability among associate degree, diploma, baccalaureate and master's-prepared nurses. *Journal of Nursing Education, 26*(9), 354-361.

Pee, B., Woodman, T., Fry, H., & Davenport, E. (2000). Practice-based learning: Views on the development of a reflective learning tool. *Medical Education, 34*, 754-761.

Pee, B., Woodman, T., Fry, H., & Davenport, E. (2002). Appraising and assessing reflection in students' writing on a structured worksheet. *Medical Education, 36*(6), 575-585.

Pesut, D.J., & Herman, J. (1998). OPT: Transformation of nursing process for contemporary practice. *Nursing Outlook, 46*(1), 29-36

Pierson, W.(1998). Reflection and nursing education. *Journal of Advanced Nursing, 27*, 165-170.

Platzer, J., Blake, D., & Ashford, D. (2000). Barriers to learning form reflection: a study of the use of groupwork with post-registration nurses. *Journal of Advanced Nursing, 31*(5), 1001-1008.

Polit. D.F. & Beck, C.T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing and Health, 29*, 489-497.

Profetto-McGrath, J. (2003). The relationship of critical thinking skills and critical thinking dispositions of baccalaureate nursing students. *Journal of Advanced Nursing, 43*(6), 569-577.

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.

Rane-Szostak, D., & Robertson, J. F. (1996). Issues in measuring critical thinking: Meeting the challenge. *Journal of Nursing Education, 35*(1), 5-11.

Rauen, C. A. (2004). Simulation as a teaching strategy for nursing education and orientation in cardiac surgery. *Critical Care Nurse, 24*(3), 46-51.

Reishman, R.R., & Yarandi, H.N. (2002). Critical care cardiovascular nurse expert and novice diagnostic cue utilization. *Journal of Advanced Nursing, 39*(1), 24-34.

Riddell, T. (2007). Critical assumptions: Thinking critically about critical thinking. Journal of Nursing Education, 46(3), 121-126.

Ruth-Sahd, L.A. (2003). Reflective practice: A critical analysis of data-based studies and implications for nursing education. *Journal of Nursing Education 42*(11), 488-497.

Sanders, D., & Welk, D. S. (2005). Strategies to scaffold student learning: Applying Vygotsky's zone of proximal development. *Nurse Educator, 30*(5), 203-206.

Scheffer, B. K., & Rubenfeld, M. G. (2000). A consensus statement on critical thinking in nursing. *Journal of Nursing Education*, 39(8), 352-359.

Scheffer, B.K., & Rubenfeld, M.G.(2007). Critical thinking: A tool in search of a job. *Journal of Nursing Education, 45*(6), 195-196.

Schleutermann, J. A., Holzemer, W. L., & Farrand, L. L. (1983). An evaluation of paper-and-pencil and computer-assisted simulations. Journal of Nursing Education, 22(8), 315-323.

Schon, D. A. (1983). *The reflective practitioner: How professionals think in action*. U.S.: Basic Books.

Sedlak, C. A. (1997). Critical thinking of beginning baccalaureate nursing students during the first clinical nursing course. *Journal of Nursing Education, 36*(1), 11-18.

Seldomridge, E.A. (1997). Faculty and student confidence in their clinical judgment. *Nurse Educator*, *22*(5), 6-8.

Shin, K. R. (1998). Critical thinking ability and clinical decision-making skills among senior nursing students in associate and baccalaureate programmes in Korea. *Journal of Advanced Nursing*, *27*, 414-418.

Shin, K., Jung, D.Y., Shin, S., & Kim, M.S. (2006). Critical thinking dispositions and skills of senior nursing students in associate, baccalaureate and RN-to-BSN programs. *Journal of Nursing Education, 45*(6), 233-237.

Simmons, B., Lanuza, D., Fonteyn, M., Hicks, F., & Holm, K. (2003). Clinical reasoning in experienced nurses. *Western Journal of Nursing Research*, *25*(6), 701-719.

Smith, A. (1998). Learning about reflection. *Journal of Advanced Nursing, 28*(4), 891-898.

Spelic, S. S., Parsons, M., Hercinger, M., Andrews, A., Parks, J., & Norris, J. (2001). Evaluation of critical thinking outcomes of a BSN program. *Holistic Nursing Practice, 15*(3), 27-34.

Spunt, D., Foster, D., & Adams, K. (2004). Mock code: A clinical simulation model. *Nurse Educator, 29*(5), 192-194.

Staib, S. (2003). Teaching and measuring critical thinking. *Journal of Nursing Education, 42*(11), 498-507.

Stevens, D.D., & Levi, A.J. (2005). *Introduction to Rubrics: An assessment tool to save grading time, convey effective feedback, and promote student learning*. Sterling, VI: Stylus.

Steward, J.L., Lynn, M.R., & Mishel, M. H. (2005). Evaluating content validity for children's self-report instruments using children as content experts. *Nursing Research, 54*(6), 414-

418.

Stewart, S., & Richardson, B. (2000). Reflection and its place in the curriculum on an undergraduate course: should it be assessed? *Assessment & Evaluation in Higher Education*, *25*(4), 369-380.

Stockburger, D. (2001). *Introductory statistics: Concepts, models and applications* (2nd ed.). Cincinnati: Atomic Dog.

Tabak, N., Bar-Tal, Y., & Cohen-Mansfield, J. (1996). Clinical decision making of experienced and novice nurses. *Western Journal of Nursing Research*, *18*(5), 534-547.

Tanner, C.A. (1987). Teaching clinical judgment. In J.J. Fitzpatrick & R.L. Tuanton (Eds.), *Annual Review of Nursing Research*, (pp. 153-173). New York: Springer.

Tanner, C.A. (1998). Clinical judgment and evidence-based practice: Conclusions and controversies. *Communicating Nursing Research, 31*, 19-35.

Tanner, C. A. (2000). Clinical judgment in nursing practice. *International Nursing Review (Japan)*, *23*(4), 66-77.

Tanner, C.A. (2006). Thinking like a nurse: A research-based model of clinical judgment in nursing. *Journal of Nursing Education*, *46*(6), 204-211.

Tanner, C.A. (2007, April). *Nursing Education: Current themes, puzzles and paradoxes*. Paper presented at the meeting of the Western Institute of Nursing in Portland, OR.

Thiele, J. E., Holloway, J., Murphy, D., Pendarvis, J., & Stucky, M. (1991). Perceived and actual decision making by novice baccalaureate students. *Western Journal of Nursing Research*, *13*(5), 616-626.

Thompson, C., & Rebeschi, L. M. (1999). Critical thinking skills of baccalaureate nursing students at program entry and exit. *Nursing & Health Care Perspectives*, *20*(5), 248-252.

Tschikota, S. (1993). The clinical decision making processes of student nurses. *Journal of Nursing Education, 32*(9), 389-398.

Vandrey, C. I., & Whitman, K. M. (2001). Simulator training for novice critical care nurses. *American Journal of Nursing, 101*(9), 24GG-24LL.

Verhonick, P. J., Nichols, G. A., Glor, B. A. K., & McCarthy, R. T. (1968). I came, I saw, I responded: Nursing observation and action survey. *Nursing Research, 17*(1), 38-44.

Videbeck, S.L. (1997). Critical thinking: Prevailing practice in baccalaureate schools of nursing. *Journal of Nursing Education, 36*(1), 5-10.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Wagner, Z. M. (1999). Using student journals for course evaluation. *Assessment & Evaluation in Higher Education, 24*(3), 261-272.

Wallace, D. (1996). Using reflective diaries to assess students. *Nursing Standard, 10*(36), 44-47.

Walsh, C.M., & Seldomridge, L.A. (2006). Critical thinking: Back to square two. *Journal of Nursing Education, 45*(6), 212-219.

Waltz, C. F., Strickland, O. L., & Lenz, E. R. (2005). *Measurement in Nursing and Health Research* (3rd. edition ed.). New York: Springer.

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.

Weller, J. M., Robinson, B. J., Jolly, B., Watterson, L. M., Joseph, M., Bajenov, S., et al. (2005). Psychometric characteristics of simulation-based assessment in anaesthesia and accuracy of self-assessed scores. *Anaesthesia, 60*(3), 245-250.

Weimer, M. (2002). *Learner-centered teaching*. San Francisco: Jossey-Bass.

Wiggins, G. (1998). *Educative Assessment: Designing assessments to inform and improve student performance*. San Francisco: Jossey-Bass.

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

Woehr, D.J., & Huffcutt, A.I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67*, 189-205.

Wooley, A.S. (1977). The long and tortured history of clinical evaluation. *Nursing Outlook, 25*, 308-315.

Wong, F. K. Y., Kember, D., Chung, L. Y. F., & Yan, L. (1995). Assessing the level of student reflection from reflective journals. *Journal of Advanced Nursing, 22*(1), 48-57.