1-1-2010

**Dependability and accuracy of clinical performance in nursing examination scores**

Laurie Vanessa Nagelsmith  
*University at Albany, State University of New York, lnagel@excelsior.edu*

The University at Albany community has made this article openly available. **Please share** how this access benefits you.

Follow this and additional works at: [https://scholarsarchive.library.albany.edu/legacy-etd](https://scholarsarchive.library.albany.edu/legacy-etd)

Part of the Educational Assessment, Evaluation, and Research Commons, and the Nursing Commons

---

**Recommended Citation**

[https://scholarsarchive.library.albany.edu/legacy-etd/225](https://scholarsarchive.library.albany.edu/legacy-etd/225)

This Dissertation is brought to you for free and open access by the The Graduate School at Scholars Archive. It has been accepted for inclusion in Legacy Theses & Dissertations (2009 - 2024) by an authorized administrator of Scholars Archive.  
Please see Terms of Use. For more information, please contact scholarsarchive@albany.edu.
DEPENDABILITY AND ACCURACY OF 
CLINICAL PERFORMANCE IN NURSING EXAMINATION SCORES

by

Laurie Vanessa Nagelsmith

A Dissertation
Submitted to the University at Albany, State University of New York
in Partial Fulfillment of
the Requirements for the Degree of
Doctor of Philosophy
School of Education
Department of Educational Psychology
2010

Author Note
This research was supported in part by a grant from the Tau Kappa Chapter of Sigma Theta Tau International Honor Society of Nursing. Correspondence concerning this dissertation should be addressed to Laurie Nagelsmith, Assistant Dean, Excelsior College School of Nursing, 7 Columbia Circle, Albany, NY 12203.
E-mail: lnagel@excelsior.edu
Abstract

The purpose of this study is to determine the dependability and accuracy of Clinical Performance in Nursing Examination (CPNE®) scores obtained from trained raters. Intraclass correlations were calculated to estimate interrater agreement. Accuracy of scores was determined by identifying percent accuracy for each scored element. Generalizability (G) coefficients were calculated to estimate sources of variance in CPNE scores. The population for the study is adjunct faculty at Excelsior College who agreed to participate in the study and are clinical examiners trained as raters for the CPNE ($N = 289$). A response rate of 25% was achieved. Each participant was asked to score the video-taped, simulated performance of two CPNE candidates using a scoring tool designed specifically for the study, which was based on the actual scoring tool for the examination with an additional section for demographic data.

Findings reveal that there is high scoring consistency or homogeneity among raters. Also, out of 105 critical elements, only 10 were scored with 97% accuracy or less. Although these accuracy findings are moderately high, given the criterion-referenced and high-stakes nature of the examination, opportunities exist for interventions aimed at enhancing the scoring accuracy of raters. Finally, application of generalizability theory for data analysis revealed that CPNE scores are dependable estimates. Small main effects were identified for candidate and rater facets with an appreciable source of variance identified for items-nested-within-candidates and residual random effects. Findings from this study add to the existing body of validity evidence for inferences made based on CPNE scores.
ACKNOWLEDGEMENTS

There are a great many people to whom I would like to express my sincere appreciation for their assistance and guidance with this research endeavor. First and foremost, I’d like thank Dr. Heidi Andrade, my committee chair, as well as the other members of my committee, Dr. Patrick Jones and Dr. Robert McMorris. I am also extremely grateful for the assistance and mentoring of Dr. Maurice Odondi during the data analysis phase of this project. Without his generosity in sharing his expertise as a statistician and psychometrician and his patience in answering my many questions, this research would not have been possible. I learned a great deal from each of these people as I persevered in refining my knowledge and skill as a researcher.

I also owe many thanks to the Excelsior College nursing faculty who shared their expertise about the Clinical Performance in Nursing Examination and helped to create the video simulation. Specifically, I’d like to thank Kathie Doyle, MS, RN who kept a “can do” attitude as we overcame challenges throughout the data gathering process and Dr. Bridget Nettleton, Dean of the Excelsior College Nursing Program, who values and supports scholarship in every way possible and has been a mentor to me for almost 10 years. Finally, I’d like to thank my family and friends who have sacrificed much throughout my doctoral education and to whom I owe a great many debts.
Table of Contents

List of Figures ......................................................................................................................... v

List of Tables .......................................................................................................................... vi

Chapter One: Statement of the Problem .................................................................................. 1

Description of the Clinical Performance in Nursing Examination (CPNE®) .................. 4
Interpretive Argument ............................................................................................................. 9
Problem Statement ................................................................................................................ 11

Chapter Two: Review of the Literature ................................................................................ 13

Clinical Competence ............................................................................................................. 13
Competency-Based and Performance Assessment ............................................................... 15
Psychometric Theory: Reliability Estimates as a Component of Validity Evidence ... 24
Psychometric Theory and Assessment of Clinical Competence in Nursing ............. 27
Research Questions ............................................................................................................. 31

Chapter Three: Methodology ............................................................................................... 32

Research Design .................................................................................................................... 32
Phases .................................................................................................................................... 35
Instruments ............................................................................................................................ 36
Participants ............................................................................................................................. 36
Procedures ............................................................................................................................. 38
Data Analysis ........................................................................................................................ 39
Accuracy ................................................................................................................................. 39
Reliability ............................................................................................................................... 40
Generalizability theory ......................................................................................................... 41
Sample size and power ......................................................................................................... 43

Chapter Four: Results ......................................................................................................... 45

Characteristics and Representativeness of the Sample .................................................. 46
Research Question 1: To what extent is there interrater agreement among raters in
scoring candidate performance during the CPNE? ......................................................... 49
Research Questions 2: To what extent are raters accurate in scoring candidate
performance during the CPNE? ..................................................................................... 49
Research Question 3: What are the major sources of variance in CPNE scores? .... 54
Chapter Five: Conclusion and Implications................................................................. 62
  Summary of Findings................................................................................................. 62
  Limitations and Delimitations.................................................................................. 67
  Implications................................................................................................................ 69
  Conclusions and Future Research............................................................................ 70

References.................................................................................................................... 72

Appendix A: Pilot #1: Measurement of the Fidelity of the Simulation Video .......... 82
Appendix B: Demographic Data and Scoring Tool..................................................... 85
Appendix C: Data Analysis with Outlier Included: Candidate Two ......................... 101
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>McCallin’s unified concept of validity model for test administration</td>
<td>26</td>
</tr>
<tr>
<td>2.</td>
<td>Percent Accuracy of Rater Scoring: Candidate One</td>
<td>50</td>
</tr>
<tr>
<td>3.</td>
<td>Percent Accuracy of Rater Scoring: Candidate Two</td>
<td>50</td>
</tr>
</tbody>
</table>
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Gender</td>
<td>46</td>
</tr>
<tr>
<td>2. RPAC Affiliation</td>
<td>46</td>
</tr>
<tr>
<td>3. Number of Months in the Rater Role</td>
<td>47</td>
</tr>
<tr>
<td>4. Race/Ethnicity of Raters in the Sample</td>
<td>47</td>
</tr>
<tr>
<td>5. Full-time Work Role of Raters in the Sample</td>
<td>48</td>
</tr>
<tr>
<td>6. Number of Times Raters in the Sample Worked in the Past 12 Months</td>
<td>48</td>
</tr>
<tr>
<td>7. Percent of Critical Elements Accurately Scored</td>
<td>51</td>
</tr>
<tr>
<td>8. Percent Accuracy for Scored Elements and Nature of Scoring Errors</td>
<td>52</td>
</tr>
<tr>
<td>9. Accuracy of Raters per RPAC</td>
<td>54</td>
</tr>
<tr>
<td>10. ANCOVA for Rater Accuracy in Scoring Candidate One</td>
<td>55</td>
</tr>
<tr>
<td>11. ANCOVA for Rater Accuracy in Scoring Candidate Two</td>
<td>56</td>
</tr>
<tr>
<td>12. Generalizability Study Results for All Areas of Care Items</td>
<td>58</td>
</tr>
<tr>
<td>13. Generalizability Study Results for All Critical Elements</td>
<td>60</td>
</tr>
</tbody>
</table>
Chapter One: Statement of the Problem

Newly licensed nurses are expected to care for patients with increasingly complex deviations from health in increasingly complex healthcare environments (Benner, Sutphen, Leonard & Day, 2010; Hamilton et al., 2007). Assuring that new graduates are qualified candidates for licensure and fit to practice the art and science of nursing is the responsibility of faculty within pre-licensure nursing education programs. An important aspect of fitness to practice is having the required clinical competencies to safely and effectively care for patients independently at a beginning-practitioner level (Rutkowski, 2007; Watson, Stimpson, Topping, & Porock, 2002). In an era of high expectations for rigor, accountability, and transparency in assessment and evaluation, it is imperative that all education programs with a clinical component incorporate sound means for the evaluation of clinical competence into their pre-licensure nursing curricula. It is also important to clearly articulate the psychometric properties of measures from which high-stakes decisions, such as ability to practice, are made (Hamilton et al., 2007; Hutchinson, Aitken, & Hayes, 2002).

Yet, despite the high-stakes nature of decisions that are made based on evaluations of clinical competence, rating the clinical competence of nursing professionals is most often exclusively left to tradition, without substantiated evidence on which to make sound interpretations from the measures used (Munoz, O’Bryne, Pugsley, & Austin, 2005; Watson et al., 2002). Although there are some merits to traditional approaches to evaluating clinical competence, such as subjective judgment and formative approaches used in apprenticeship-type educational models, often they are insufficient for making necessary justifiable inferences in relation to clinical competence (Benner, et al.,
2010; Cox, 2000; Lenburg, 1976; 1979a; 1979b). That is, such approaches do not provide sufficient validity evidence on which to make sound inferences from assessment scores (Pulito, Donnelly, & Pylmale, 2007).

In addition to the risk of making erroneous decisions about the achievement of specified educational program outcomes, candidacy for licensure examinations, and subsequent fitness to practice autonomously in clinical settings, drawing inferences from measures of clinical competence that are not psychometrically sound and transparent has legal, policy, and accreditation consequences (Linn, 1993, 2006). Lack of available validity evidence for the use of test scores makes legal defensibility tenuous, has the potential to lead to poor policy decisions, and calls into question a program’s ability to meet professional accreditation standards (Lang & Wilkerson, 2008; Rutkowksi, 2007; Schuwirth et al., 2002; Schuwirth & van der Vleuten, 2004). For these reasons, providing sufficient validity evidence on which to draw sound inferences from the scores of clinical competency measures should be a primary focus of all education programs with a clinical component, such as programs that prepare graduates for licensure as registered nurses (RN).  

Stakeholders in assuring clinical competence of newly graduated health care providers, such as RNs, include consumers of health care services, others actively practicing in the profession, other members of the healthcare workforce, educators, institutions of higher learning, state regulators, and legislators. These stakeholders have a strong vested interest in assuring that newly graduated nurses have the full complement of knowledge, attitudes, and skills to practice nursing safely and effectively (Rutkowski, 2007; Schuwirth et al., 2002). However, despite agreement on the importance of
measuring clinical competence in a psychometrically sound manner, many charged with overseeing the education and practice of health care providers do not pay sufficient attention to sound measurement of the outcomes of clinical learning (Downing, 2004; McNeil, Hughes, Toohey, & Dowton, 2006; National League for Nursing, 2008; Watson, 2000; Watson et al., 2002). All too often, decisions about candidacy for RN licensure are based on a limited view of the process of nursing education, many times at the exclusion of the more important issues around the achievement of learning outcomes and the psychometric integrity of how outcomes of education are measured (Benner et al., 2010; National Council of State Boards of Nursing [NCSBN], 2005; 2006a; 2006b; 2006c; 2006d).

This limited perspective on nursing education is depicted in the literature as a longstanding debate about the nature and effectiveness of competency-based education (Lenburg & Mitchell, 1991; Schmidt & Lyons, 1969; Smith & Lichtveld, 2007; Spady, 1978; Tanner, 2001; Woodward, 1976). To engage in a meaningful debate, it is important that nursing educators (a) understand the factors involved in the evaluation of clinical competency and the importance of creating and maintaining measures of clinical competence that provide sufficient levels of validity evidence to justify decisions based on evaluation outcomes, (b) diligently and continuously pursue the systematic collection and analysis of validity evidence, and (c) assure that validity evidence and threats are fully transparent to stakeholders. In a study sponsored by the Carnegie Foundation, Benner et al. (2010) found evidence that most nursing curricula are fragmented with a lot of information being taught without sufficient efforts toward finding out whether students understand the information. Thus, one of the recommendations for transforming
nursing education made by these researchers is requiring performance assessment for RN licensure.

Characteristics of measurement that support or threaten the psychometric integrity of an assessment process directly impact the validity of inferences based on examination scores (Downing, 2006; Kane, 2006; McCallin, 2006; Pulito, Donnelly, & Plymale, 2007). Considerable research has focused on characteristics of measurement in standardized achievement tests in K-12 education and tests of theoretical knowledge in post-secondary as well as post-licensure professional certification (Hutchinson et al., 2002; Linn & Gronlund, 2000; National Research Council, 2001). Factors related to assuring acceptable levels of validity evidence in the assessment of clinical competence have not been sufficiently researched (Cox, 2000; Hays et al., 2002; Munoz, O’Bryne, Pugsley, & Austin, 2005). For a variety of reasons, there is a paucity of research focusing on assessment of learning outcomes in clinical education, particularly with respect to validity evidence linked to measurement scores on which critical decisions are made. Likewise, factors related to assuring sufficient validity evidence for making decisions about clinical competence, including interrater reliability, have not been sufficiently studied (Defloor et al., 2006; Watson et al., 2002).

**Description of the Clinical Performance in Nursing Examination (CPNE®)**

Evaluating and reporting validity evidence and threats that impact inferences based upon examination scores is vital for assuring psychometric integrity. The precision and dependability of measurement for a well-established measure of clinical competence, the Clinical Performance of Nursing Examination (CPNE), is the focus of this study.
The purpose of the CPNE is to serve as a summative, capstone assessment for Excelsior College’s associate degree in nursing program. Excelsior College was originated as the Regents External Degree Program of the University of the State of New York and later evolved to become Regents College in 1971. In 2000, Regents College was granted a charter as an independent institution of higher education and the college name was changed in 2002 to Excelsior College.

The CPNE is the clinical capstone examination for Excelsior College’s competency-based, associate degree in nursing program developed, piloted, and first administered in the early 1970s (Lenburg, 1979a). It is a mature measurement tool with an over-30-year history of use in pre-licensure nursing education. The faculty committee tasked with constructing the original CPNE adhered to a systematic approach to performance examination development using a model that focused the faculty on examination content, examination characteristics, and the structure and process for administration and scoring (Yarbrough, Jones, Nagelsmith, Kim, Snead, & Odondi, 2007). The original CPNE creation process adhered to concepts in more contemporary models for test administration, which are compatible with current standards for test development (e.g., McCallin’s Unified Concept of Validity Model for Test Administration [McCallin, 2006]). For example, concept mapping and development of a test blueprint took considerable time and effort and ultimately led to a performance examination with an acceptable level of content validity determined through expert faculty review and convergence with factors such as pass rates on the national licensing examination and supervisor surveys on graduate performance.
Ongoing evaluation of content validity of the CPNE is assured through a process whereby a nursing faculty committee consisting of practitioners and educators from across geographic areas and nursing specialties continuously monitor and refine examination content as well as examination administration and scoring protocols. Modifications to the examination are made by this group of expert nurses based on new evidence in nursing science and other factors that drive practice changes. Also, the examination is periodically reviewed by psychometricians and subject matter experts to ensure that acceptable psychometric technical properties are achieved.

The CPNE is a dichotomously scored, criterion-referenced, performance examination administered over two and one-half days at 20 hospital test sites throughout the United States. It is a complex performance examination that takes place in the authentic patient care environment with actual patients. Faculty who administer the examination, including observing and scoring candidate responses to test stimuli, function as raters and are called clinical examiners. The faculty rater-to-candidate ratio during the examination is one-to-one. Tasks related to clinical decision making and application of each element of the nursing process (i.e., assessment, diagnosis, planning, implementation, and evaluation) are scored throughout the examination, and the results are interpreted as a measure of competence in critical thinking.

The examination covers the full domain of knowledge and skill required for beginning level nursing practice through application of a systematized sampling methodology used throughout the examination process. The sampling frame of the examination is organized into areas of care, which are operationalized as critical elements. Critical elements are single, discrete, observable behaviors that reflect an
aspect of competence for an area of care. Critical elements reflect essential nursing actions that are based on principles and theories relevant to the art and science of nursing (Lenburg, 1979a). They are not merely rote atomistic behaviors, but rather are distinct nursing actions that require integration and synthesis of complex nursing knowledge applied to actual patient care needs. The examination sampling frame consists of 600 critical elements from which more than 200 critical elements are sampled during each CPNE administration (Yarbrough et al., 2007). Examples of critical elements related to the area of care for vital signs are (1) obtains accurate vital signs by reading the instrument within a stated range of +/-6 millimeters for blood pressure, and (2) obtains oxygen saturation when assigned.

Systemized conditions in CPNE structure and process are attained through standardized examination development, administration, and scoring procedures. There are two major components of the examination structure; the nursing simulation laboratory (NSL) component and the patient care situation (PCS) component. To pass the examination, candidates must demonstrate competence with 100% accuracy in each of the NSL stations, in two of a possible three adult PCSs, and one of a possible two pediatric PCSs.

These defined standards for acceptable performance to pass the examination were established by the original CPNE developers, who in concert with expert clinicians, determined exactly what level of performance is required of newly graduated nurses. Standards for passing that require 100% accuracy in performance of the NSLs and PCSs were established to allow for the inferences made from CPNE scores. That is, passing the examination along with all other nursing degree requirements provides a sound basis to
support the conclusion that an Excelsior graduate has rigorously demonstrated all relevant program outcomes required for the associate degree in nursing, and is therefore eligible to seek RN licensure. Tolerance for error by allowing the candidate the ability to repeat NSL stations once and an adult and pediatric PCS once is congruent with standards for setting cut scores and acknowledges unforeseen and uncontrollable issues that may arise when administering a high-stakes performance examination in the authentic patient care environment. Such issues may be due to lack of adherence to standardized examination administration protocols for reasons such as changes in the patient’s condition and thus the plan of care during the PCS, or to deviations from standardized scoring criteria by raters for reasons such as insufficient training or ambiguous scoring criteria. Regardless of the causes of deviations from systematized conditions and standardized scoring, the bottom line result is the introduction of error into the candidate’s examination score.

The NSL component consists of four simulation stations during which candidates must demonstrate competency in wound protection, intravenous (IV) medication administration, injectable medications/IV push administration, and injectable medications/intramuscular and subcutaneous administration. Sampling competencies is not used for the NSL portion of the examination because the faculty believes the competencies addressed in the NSL are so important that they must be successfully demonstrated by all candidates. During the NSL component, candidates must demonstrate a total of 41 critical elements with 100% accuracy.

The PCS component of the CPNE involves a more complex administration methodology because of the inherent complexity of testing clinical competency in the
authentic patient care environment. Each PCS is conceptualized as a distinct test item. Test stimuli are operationalized as areas of care. Examples of areas of care are vital signs, fluid management, asepsis, caring, wound management, and abdominal assessment. Comparability is maintained by controlling the number of areas of care assigned to each candidate for each test item and by adhering to examination administration protocols which specify criteria for patient selection by the clinical examiner (i.e., the rater).

Areas of care are categorized as “overriding”, “required”, or “selected”. The overriding areas of care are deemed critically essential and relevant for safe and effective care in all aspects of nursing practice. Therefore, scoring criteria in relation to these areas of care are in effect during each PCS. Overriding areas of care include asepsis, caring, emotional jeopardy, physical jeopardy, and mobility. Other areas of care, called required areas of care, are considered so important to nursing competence that they are tested at least once per PCS. Required areas of care include application of the nursing process, fluid management, medications, and vital signs. The remaining areas of care are termed selected areas of care. It is the selected areas of care that are randomly sampled in a manner that is intended to support validity evidence for inferences derived from CPNE scores (Yarbrough et al., 2007). During the PCS component of the examination, a candidate must demonstrate all critical elements for assigned selected and required areas of care as well as all critical elements for the overriding areas of care.

**Interpretive Argument**

Yarbrough et al. (2007) clearly define the interpretive argument for use of examination scores:
Successful achievement of clinical learning outcomes measured by the CPNE®, coupled with learning outcomes met through successful completion of all other associate degree in nursing examinations and general education requirements, provide a sound basis to support the conclusion that an Excelsior College graduate has rigorously demonstrated all relevant program outcomes required for the associate degree in nursing, and is therefore eligible to seek RN licensure and subsequently practice as a registered nurse. (p. 8)

The interpretive argument is theoretically based on a unified approach to validity (McCallin, 2006; Messick, 1989), assumptions underlying competency-based education and the nature of performance assessment, and measures of divergent and convergent validity. The interpretive argument refers to valid inferences that can be made about the scores of the examination for their intended use. Such inferences are based on assurance of sufficient validity evidence. Validity evidence includes, but is not limited to, confirmation of careful test construction and ongoing analysis of the psychometric properties of the examination (McCallin, 2006; Messick, 1989). Evaluation of validity evidence and threats were recently reported in the technical manual for the CPNE, which reveal sound evidence on which to base the above-stated interpretive argument (Yarbrough et al., 2007).

The validity evidence for the CPNE has historically been grounded in a unified approach to validity, which involves attention to many factors that relate to construct-irrelevant variance (CIV) and construct under representation (CUR) (McCallin, 2006). Because validity refers to the ability of an instrument to measure what the assessor intends to measure, it is imperative to minimize CIV and assure that all relevant constructs are represented and adequately sampled by the measure (Lang & Wilkerson, 2008; Messick, 1989).
One means to minimize CIV is to ensure consistency and precision of measurement to the extent possible (McCallin, 2006). The precision of an instrument is usually evaluated based on the reliability of scores obtained from the administration of the instrument, and it is an important characteristic of any examination (Anastasi & Urbina, 1997). Reliability of examination scores is a necessary but not sufficient condition for the validity of inferences associated with an interpretation of the exam scores (Downing, 2004; Polito, Donnelly, & Plymale, 2007). Scores showing certain characteristics lend credence to inferences suggesting that a test produces scores which show evidence of reliability. Therefore, it is important to establish measures of consistency and accuracy of the raters scoring candidate performance in the clinical environment as well as to determine the dependability of scores by partitioning and estimating the major sources of variance in scores, which are the foci of this study. This additional analysis of the accuracy and dependability of examination scores will build on evidence from a study conducted by Excelsior College faculty and staff in 2005 that was grounded in both a unified construct of validity model and generalizability theory (Brennan, 2001; Gao & Brennan, 2001; Shavelson & Webb, 1991; Yarbrough et al., 2007).

**Problem Statement**

The present study will add to the body of evidence regarding the interpretive argument for decisions based on CPNE scores. Specifically, accuracy of the scoring of the CPNE by trained raters as well as measures of consistency in scoring based on classical test theory, including estimates of interrater reliability, will be determined. In
addition, dependability of CPNE scores will be explored by applying generalizability theory to estimate major sources of variation in scores.
Chapter Two: Review of the Literature

This study is grounded in the theoretical underpinnings of clinical competence, competency-based and performance assessment, and psychometric theory, including Messick’s (1989) multifaceted concept of validity and McCallin’s (2006) Unified Concept of Validity Model for test administration. An examination of empirical evidence relevant to the present study involves a review of the research related to clinical competency assessment, existing validity evidence for the interpretation of CPNE scores, and evaluation of interrater reliability as it relates to competency and performance assessment. A search of the literature from 1969 to the present was conducted using the search databases within EBSCO Host, including (a) Academic Search Premier, (b) CINHAL Plus, (c) ERIC, (d) PsycArticles, (e) MEDLINE, and (f) Professional Development Collection.

Clinical Competence

There are controversies and challenges in defining and measuring clinical competence in health care disciplines. Many have said that the first step to addressing clinical competence is to clearly define it (Tilley, 2008; Hamilton et al., 2007; Nagelsmith, 1995; Watson et al., 2002). On the surface this appears a straightforward step; however, there is little consensus in the literature about the nature and constituent parts of clinical competence (Hays et al., 2002; Rutkowski, 2007). The discord between opposing views of clinical competence is based on disagreement about whether competence refers to actual or potential ability, as well as diverging opinions about the differences and similarities between the concepts of competence and performance (Hager & Gonczi, 1994; Schuwirth et al., 2002; Watson, 2000; Watson et al., 2002). For
example, there are situations when a broad view of competence is useful and appropriate, such as this author’s conceptualization of competence in relation to staff development, which is defined as, “an unending potential for acquiring knowledge and skill through critical thinking experiences and an attitude of openness to life and learning” (Nagelsmith, 1995, p. 246). Other situations, such as evaluating pre-licensure clinical competencies, require a definition of competence in relation to specific standards of practice (Hamilton et al., 2007; Smith & Lichtveld, 2007). Competency is often also defined in an integrated fashion, such as, “knowledge, abilities, skills, and attitudes displayed in the context of a carefully chosen set of realistic professional tasks which are of an appropriate level of generality” (Hager & Gonczi, 1994, p. 3).

 Watson et al. (2002) analyzed the research evidence for the use of clinical competence assessment in nursing through an integrated review of the literature published between 1980 and 2000. The researchers found few papers with sufficient rigor to support an intellectual stance about how to clearly define and measure clinical competence in a psychometrically sound manner. Ultimately the authors were overwhelmed by the multitude of definitions and contradictions in the literature and concluded, “. . . it would appear that the search for competence is a bottomless pit” (p. 428).

 Despite the despair of these researchers, it seems likely that considering clinical competence within the context of established standards of practice and based on expert judgment allows educators the ability to climb out of the pit by grabbing onto theoretically sound principles of adult learning and psychometrically sound methods of competency assessment and performance assessment. For the purposes of this study,
clinical competence will be defined as performance of nursing care that meets defined and predetermined standards of performance. Using this definition, clinical competence is the demonstration of clinical knowledge and skills that are measured by a trained faculty rater via a rigorous and systematized summative clinical performance assessment. That is, competence is not assumed by completion of formative clinical learning experiences coordinated by multiple faculty members without attention to the psychometric properties of the means for the measurement of competence.

**Competency-Based and Performance Assessment**

According to Hager and Gonczi (1994), competency-based assessment allows a means to determine if candidates meet predetermined standards of performance. Although competency-based assessment is not new, many educators and stakeholders are not familiar with its principles, essential properties, and uses. Lack of understanding of competency-based assessment often leads to misconceptions about the nature of this type of evaluation as well as poorly designed measures of competence (Hager & Gonczi, 1994; Lenburg, 1991). For example, a major worry among some educators and regulators is that competency-based assessment is atomistic and cannot capture the richness of competency required for professional practice. Yet, if competency-based assessments are based on established standards of practice and an integrated conceptualization of competency, and if they are designed based on sound psychometric theory, then measures can be created to capture the complex and holistic nature of beginning nursing practice. Likewise, with astute attention to requisite principles and properties, performance assessments for measuring continued competence in general or specialty nursing practice can be created and implemented.
Although some aspects of practice must be artificially fragmented to make assessment feasible, the cognitive complexity of the clinical reasoning that takes place during interactions with the patient during task performance should not be overlooked by focusing solely on the observable behaviors employed to carry out the task. Hager and Gonczi (1994) stress that emphasis on the application and synthesis of knowledge required in well-designed, integrated competency assessments, “leads ultimately to a greater integration of the theoretical and practical, which is at the heart of successful professional practice” (p. 6). Thus, it is important to clearly identify the comportment and cognitive processes one desires to assess, identify behaviors that illustrate them, and then go about the measurement process in a psychometrically sound manner according to set standards for educational testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, [AERA, APA, NCME], 1999). This focus on the three major points of clinical competency assessment, including cognition, observational models, and measurement models, is aligned with what the National Research Council (2001) refers to as the assessment triangle.

Performance assessment is a subclass of competency assessment with a long history in nursing education and other health care disciplines (Hawes, 1978; Hays et al., 2002; Hyland, 1978; Lenburg, 1976, 1979a, 1979b; Lenburg & Mitchell, 1991; Linn & Baker, 1996; McNeil et al., 2006; Nolan, 1998; Robbins & Hoke, 2008; Schmidt & Lyons, 1969; Wozniak, 1973; Yarbrough et al., 2007). Performance assessment is integral to all health care professions and is used to determine beginning clinical competency as well as ongoing clinical competency. It has been defined in a number of
ways, including, “a coherent evaluation system which assesses whole occupational functioning including its constituent parts” (Hamilton et al., 2007, p. 776) and as “the process of gathering data by systematic observation for making decisions about an individual” (Berk, 1986, p. 24).

Despite different forms, performance assessments share similar features. They are preferred mechanisms for evaluation of contextual, real-world problem solving and evaluation of higher-order thinking skills. Linn and Baker (1996) identify several characteristics of performance assessment, including the construction of a response or performance of an activity, employment of higher-order thinking skills, extensive periods of time for performance, allowance for some latitude in the choice of response, labor intensive scoring processes, and dependence on human decisions in scoring. The same authors also emphasize the importance of measures to ensure accuracy and consistency in scoring performance assessments, including unambiguous scoring criteria and rater training.

Performance assessments are selected as an evaluative mechanism for a number of reasons. One paramount reason is that they can lead to greater and more appropriate accountability. Accountability as the rationale for use of performance assessments accentuates the need for a clearly defined interpretive argument for the purpose and use of the assessment scores as well as adherence to psychometrically sound principles (Linn & Baker, 1996). Specifically, decisions made based on the results of performance assessments of clinical competency must be grounded in sufficient validity evidence and overt acknowledgement of any potential threats to validity (Linn & Baker, 1996;
McCallin, 2006; Welch, 2006). Linn and Baker state unequivocally, “technical quality is an unassailable requirement of performance assessment” (p. 100).

When creating performance assessments it is essential to manage the tension between authenticity and standardization necessary to minimize measurement error (Welch, 2006). Because of the nature of performance assessments, Linn and Baker (1996) purport that some level of unreliability will always be reflected in the scores derived from them. For this reason, it is essential to take steps to minimize error of measurement in scores to the extent possible and evaluate more problematic forms of reliability with these types of measures, such as interrater reliability (Downing, 2006). Additionally, there is consensus that increasing the number of tasks involved in performance assessments is more effective in increasing reliability in these types of measures than increasing the number of raters (Linn & Baker, 1996; Livingston & Lewis, 1993).

It is important to put the current study into the context of existing research that has been conducted on various aspects of competency-based education and the CPNE, including inferences based on CPNE scores that provide evidence of reliability and validity. There have been a number of research studies focusing on the psychometric properties of the CPNE and competency-based education. Because Excelsior College, formerly the Regents External Degree Program, and Regents College, is the only nursing program in the world that is based entirely on a competency-based model of education and includes the CPNE as the clinical capstone requirement for the program, all of the studies discussed in this review of the literature are relevant.
Hyland (1978) compared CPNE performance between graduates of the then Regent’s External Degree competency-based associate degree (AD) in nursing program with graduates of AD in nursing programs with traditional models of nursing education. The researcher hypothesized that graduates of traditional programs would perform better than the competency-based program graduates, and that areas of care that were the cause of failure would be similar for both groups. Hyland’s findings did not support either hypothesis.

Hawes (1978) compared graduates of various types of AD in nursing education programs on their transition into professional nursing practice by comparing supervisor ratings to measure nursing competencies using the Dyer Descriptive Scales for Nursing Performance. Research involved comparing ratings of graduates of traditional programs with graduates of an assessment-based, competency program (i.e., Regents College) Findings revealed no significant differences in supervisor ratings of clinical competence between graduates of a traditional and competency-based AD in nursing programs.

Later, Regents College commissioned the State University of New York at Albany to conduct a study to evaluate the quality and rigor of the competency-based AD in nursing program using a survey design with graduates and their supervisors (Newman, Stahl, Pierce, & Borelli, 1995). One thousand evaluation surveys were sent first to graduates and then to supervisors. The response rate for graduates was 32% and the response rate for supervisors was 60%. The results indicated that the demographic and prior health care experience of graduates of the competency-based program were different than graduates of traditional programs and that supervisors consistently rated Regents College graduates slightly better than other AD graduates in terms of clinical
skills, interpersonal skills, knowledge, professionalism, supervisory skills, and overall job performance. The vast majority of supervisors were satisfied or very satisfied with the performance of graduates in clinical health care settings.

Excelsior College also commissioned two independent research studies in 2009 focusing on supervisors’ perceptions of the work performance of Excelsior College graduates. One study conducted by SRI International, an independent nonprofit research institute, used a two-tiered survey design to gather data from nurse supervisors of Excelsior College’s newly licensed graduates (i.e., graduates between 2006 and 2008). One element of the survey involved asking supervisors identified by graduates to compare Excelsior graduates to newly licensed graduates of other programs. The other element involved a pre-recruited research panel of nurse supervisors from a range of healthcare facilities across the country. Both surveys used a 7-point Likert scale with items addressing clinical competency of graduates, program quality, and a series of items addressing specific nursing competencies (Gwatkin, Hancock, & Javitz, 2009). Overall, both groups surveyed reported that Excelsior College graduates performed the same as or better than traditionally-prepared graduates. A large majority of both groups reported that, if given the opportunity, they would hire an Excelsior AD in nursing graduate (96% and 80%, respectively). In addition, 92% of graduate-source supervisors and 81% of panel supervisors rated the overall quality of the Excelsior program as the same as or better than other programs (Gwatkin, Hancock, & Javitz, 2009).

The second recently conducted study was carried out by ProEvaluators, LLC. These researchers used a telephone interview and survey design to gather data. In addition, synthesis and analysis of existing data obtained from the National Council of
State Boards of Nursing and Excelsior College were used to address some of the research questions. Key findings by the researchers included that 95% of respondents rated Excelsior graduates as slightly better prepared for the workforce than traditional graduates and that given the opportunity 95% of the respondents would hire another Excelsior College graduate if given the opportunity (Darrah & Humbert, 2009).

In addition to the aforementioned studies on competency-based education in general with outcomes lending credence to the intended use of CPNE scores, there have been a number of studies focusing specifically on the psychometric properties of the CPNE. A significant study focusing on the reliability of CPNE scores was conducted by Nettleton in 1996. Nettleton’s design focused on raters, and analyzed variables influencing the accuracy and consistency of scoring the CPNE using videotaped simulation. Results suggest that raters scored the CPNE with a high degree of inter-rater reliability and individual accuracy. Of the 52 critical elements scored, 43 were scored with 90% agreement with a mean accuracy score of 88.8%. In addition, correlation coefficients for score and selected rater characteristics (i.e., age, race, highest earned educational degree, year training completed, frequency of working in the rater role, regional location, and amount of time spent in the work role with student/graduate nurses) were not statistically significant.

Nettleton’s research design differed from the present study in a number of ways. A major design difference is in the delivery of the video simulation. In Nettleton’s study, the video simulation was scored in a face-to-face manner with Nettleton traveling to Regional Performance Assessment Centers and having raters view the stimulus in cohorts in a physical location on a television screen. In the present study, the video simulation
was scored via a webcast session on a computer monitor without convening the raters in a physical location. Another difference is that raters engaged in a discussion after each scoring session so that Nettleton could debrief the group on their experience scoring the video simulation and inform the raters regarding accuracy of scoring. If there were ambiguities in the video depiction of the CPNE or if there were video related issues with scoring, these points were addressed during the discussion and considered prior to data analysis. Also, Nettleton’s design involved the scoring of one candidate’s performance on 9 areas of care and 52 critical elements, while the present design involves the scoring of two candidates’ on 15 areas of care and 105 critical elements. It is also important to note that both study designs involved voluntary participation. Nettleton’s face-to-face design resulted in a 75% response rate and the present study webcast design resulted in a 25% response rate.

In 1999, Heywood investigated the content validity of the CPNE in relation to its representativeness of the domain of required nursing knowledge and skill. Participants included a panel of experts in all specialty areas of nursing practice relevant to pre-licensure nursing education. Participants agreed that domain definitions for areas of care were complete and representative as well as appropriate for the knowledge and skills expected of an associate degree in nursing graduate.

The most recent research focusing on dependability of CPNE scores was conducted in 2005 (Yarbrough et al.). Evidence of equality of testing outcomes was established through the application of generalizability (G) theory in determining stability and dependability of the examination scores. In this analysis, G coefficients and standard errors of measurement suggested that CPNE scores are stable estimates of clinical
competence. However, there was evidence of variability in CPNE outcomes between ethnic groups and among hospitals where the examination was administered. While it was verified that these outcome differences were consistent between the sample tested and the entire population of test takers, the finding warrants further analysis (Yarbrough et al., 2007). In summary, the aforementioned studies provide validity evidence for the intended use of CPNE scores. The current study will add to the body of validity evidence for the intended use of CPNE scores and identify threats to validity.

The importance of focusing on discrete observable behaviors during performance assessment in the clinical setting as is the case with the CPNE, rather than more traditional approaches to evaluation in the clinical setting, is illustrated in a study focusing on global ratings of the clinical performance. In a study focusing on global ratings of clinical performance of medical student candidates, Pulito et al. (2007) studied faculty rating practices. This study was conducted to ascertain if some particular student performance characteristics were more highly related to the overall faculty grade than other characteristics. Researchers acknowledged at the start of the study that there is consensus in the literature that faculty tend to rate clinical performance based on one overall characteristic of the student rather than rating various components of clinical performance independently. The research design was aimed at determining if measurement of clinical performance could be determined by anything other than a general factor. This research revealed low interrater reliability for the faculty rating of medical students’ clinical competence and determined that rating on any one performance characteristic had a predictive validity of 75-80% for a student’s overall grade. Interestingly, faculty scored criteria related to personality characteristics of students the
highest and scored criteria related to cognitive skills the lowest. Ultimately, the researchers concluded that faculty tended to make simplistic, undifferentiated determinations of the clinical performance of students. These findings suggest that scoring protocols and tools restricted to discrete observable behaviors that reflect synthesis and integration of complex cognitive processes would be one means to enhance the validity of scores derived from measures of clinical performance.

The studies reviewed here indicate a substantive body of evidence for the intended use of the CPNE scores. However, because the CPNE is periodically refined to reflect contemporary nursing practice and because new raters are frequently trained and incorporated into the existing mix of raters, there is a need for additional evidence to ensure that clinical performance is measured accurately and consistently. In addition, reinforcing existing research that further elucidates the dependability of CPNE scores, including components of score variance, will add to the body of evidence regarding the interpretive argument for decisions based on scores.

**Psychometric Theory: Reliability Estimates as a Component of Validity Evidence**

Validity is, “the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of the tests” (AERA, APA, NCME, 1999, p. 9). Messick (1989) identified a unified, multifaceted approach to validity that establishes lines of validity evidence rather than speaking of various types of validity. Using this integrated approach to validity encompasses reliability as a form of validity evidence. Reliability addresses the consistency and precision of assessment results. Ultimately, the line of validity evidence, including estimates of reliability, must substantiate the meaning
of an examination. Creators and users of examinations must provide sufficient evidence
to ensure a sound argument for interpretation of examination scores. Kane (2006) goes
further to explain that any validity argument is only as strong as its weakest link meaning
that if an examination is problematic in terms of any line of validity evidence, then the
interpretive argument for use the of the examination scores or measurement outcomes is
not plausible. Another important point to consider when evaluating validity evidence,
including data related to reliability, is sources of variance in the examination scores.
Sources of variance in scores can be determined and analyzed via application of
generalizability theory. Identifying sources of variance is important because less
certainty could be placed in examination scores if more error variance were reported as
part of a candidate’s score.

McCallin’s unified concept of validity model for test administration (2006)
depicted in Figure 1 identifies two types of errors that can occur during test
administration, (1) construct underrepresentation, and (2) construct-irrelevant variance.
Construct underrepresentation results in narrowing the meaning of examination scores. It
occurs when examinations are poorly constructed, such as when the entire domain of
interest is not reflected in the universe of knowledge from which examination items are
sampled or an adequate sampling plan is not used. Construct-irrelevant variance results
from systematic error. It occurs when examination scores are affected by processes or
factors unrelated to the construct of interest being examined.
There is clear agreement that no measure captures the entire domain of interest and no score or rating is without error (Anastasi & Urbina, 1997; Brennan & Kane, 1979; Buckendahl & Plake, 2006; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Gall, Borg, & Gall, 1996; Livingston & Lewis, 1995; Nunnally, 1978; Thorndike, 2005). The questions always are how to minimize threats to validity of examination score interpretations, such as how to minimize construct-irrelevant variance and how to assure that all relevant constructs from the domain of interest are included in the assessment.
(McCallin, 2006). Examples of ways to minimize construct underrepresentation include careful delineation of all elements of the domain of knowledge being tested and evidence of careful test construction. Some ways to minimize construct-irrelevant variance include evidence of appropriate test administration and acceptable reliability and errors of measurement in relation to the intended use of scores.

**Psychometric Theory and Assessment of Clinical Competence in Nursing**

The final theoretical foundation for the study, in addition to clinical competence and competency-based education and performance assessment, is psychometric theory. Psychometric theory and its relevant concepts (i.e., validity and reliability) and models (i.e., unified concept of validity) need to be consistently applied when assessing clinical competence in nursing practice.

Assessment of clinical competence in traditional nursing education typically occurs in a formative fashion during a clinical rotation with one or more faculty members overseeing the learning experiences of a group of students in collaboration with identified RN preceptors or mentors in the health care environment (Watson, 2000). In this type of learning situation, evaluation occurs over time and there are not the systematized conditions and attention to the psychometric theory and principles that allow for the type of critical inferences that are made from the clinical evaluations.

Simulations are increasingly being incorporated into clinical education programs primarily for learning purposes but are also used to evaluate clinical competency in narrowly delineated areas of practice. Objective structured clinical examinations (OSCE), a specific type of simulation, are also recognized as means for assessing clinical
competency for defined constructs. Many times OSCEs use some form of technology to present the candidate with stimulus to respond to. Sometimes OSCEs use simulated live patients programmed to present to the candidate with reports of particular deviations from health that should elicit particular nursing responses (Munoz et al., 2005). Simulations allow for the type of systematized conditions necessary for psychometrically sound examination of competency. Yet, to date the literature does not reflect sufficient attention to the psychometric integrity of such assessments.

Both traditional clinical rotations and simulations are useful learning strategies and means for formative assessment. However, in terms of making critical decisions about candidate performance, it could be argued that they lack a sufficient line of validity evidence upon which to make high stakes summative decisions. The CPNE is one such measure with a line of evidence that supports the intended use of CPNE scores. Validity evidence and threats are clearly depicted in the CPNE technical manual (Yarbrough et al., 2007).

The CPNE is a summative competency-based, performance assessment with documented evidence of validity to support the interpretive argument for examination scores. The examination is administered by trained faculty raters in the authentic patient care environment. Examination prompts are presented by actual patients experiencing actual deviations from health. Candidates for the CPNE are Excelsior College associate degree in nursing students who have completed all other AD in nursing program requirements and are within 10 credits of completing general education requirements.
Evidence supporting the interpretive argument for the CPNE includes observance of the steps of careful test construction identified by Downing (2006), which include adherence to the relevant *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999). Additional evidence includes substantiation of appropriate scoring (Nettleton, 1996) and evidence of attention to fairness and appropriate test administration, equality in testing outcomes, and evidence of accurate score interpretation (Yarbrough et al., 2007).

Stability and dependability of performance assessment scores are vital components of establishing estimates of reliability (Downing, 2004). The paramount importance of establishing acceptable levels of reliability is commonly cited in the psychometric literature by emphasizing that examination scores can be reliable even if they are not valid, but scores cannot be valid without being reliable (Anastasi & Urbina, 1997; Linn & Gronlund, 2000). It is therefore vitally important to build on prior psychometric research on validity evidence and threats to the interpretation of CPNE scores by periodically establishing estimates of interrater reliability because of changes in raters as well as modifications to standardized administration and scoring criteria (Downing, 2004; Yarbrough et al., 2007). Establishing levels of reliability for examination scores is one of many essential elements for identifying validity evidence and threats. Ultimately, there is a need to achieve expert ratings of nursing clinical competence through measures that control for bias and are consistent and accurate measures of the constructs of interest.

Further psychometric analysis of the CPNE will elucidate lines of validity evidence by providing supplemental support for the interpretation of scores or by
identifying threats to validity of the interpretive argument. The present study focuses on interrater reliability of scoring by clinical examiners (i.e., raters) who meet eligibility requirements and have been trained to act in the role. In addition, the extent to which raters are accurate in scoring the examination and the extent of association between various rater characteristics and scoring accuracy will be revealed. In addition to providing current evidence related to consistency and precision of scoring, findings may be more generally considered with other relevant, albeit sparse, literature focusing on the assessment of clinical learning outcomes of pre-licensure nursing candidates.

Analyzing the accuracy and dependability of measures of clinical learning outcomes in nursing education should replace the all too common practice of exclusively relying on formative expert judgment of candidate competence without substantive technical reporting of the psychometric properties of the measures used (Downing, 2006; Kane, 2006; Polito et al., 2007; Schuwirth & van der Vleuten, 2006). Such analysis is necessary, because educators must make determinations regarding the clinical competence of students using measures that have sufficient, documented validity evidence.

Achieving consistency and precision are a challenge in clinical performance assessment because of the need to assess many aspects of competence in the authentic patient care environment. In authentic environments, some patient and setting variables are outside of the control of the faculty member or candidate; assuring precision, consistency, and dependability in measurement is vital to minimizing the amount of error in measurement to the extent possible.
Research Questions

The fact that psychometric properties are not static qualities of an examination, combined with the value of providing sufficient validity evidence for the intended use of CPNE scores, including better understanding sources of variability in scoring the examination, has led to the following research questions:

- To what extent is there interrater agreement among raters in scoring candidate performance during the CPNE?
- To what extent are raters accurate in scoring candidate performance during the CPNE?
- What are the major sources of variance in CPNE scores?
Chapter Three: Methodology

The methodology for this research is intended to estimate the accuracy and consistency of scoring of the CPNE by trained raters and apply generalizability theory to identify measures of the dependability of scores. The CPNE is a criterion-referenced performance examination taken by candidates as the capstone requirement of an associate degree in nursing program.

The researcher used a digitally recorded video simulation of aspects of the CPNE as a stimulus for scoring. Specifically, the Patient Care Situation (PCS) component was the focus of analysis. Scores assigned by raters to each candidate in the video simulation were used to analyze scoring accuracy, interrater reliability, and sources of error variance, thereby estimating the precision, consistency, and dependability of scores.

Research Design

This study used a digitally recorded simulation of two candidates performing critical elements for different areas of care. Only elements of the PCS component of the CPNE were scored because the PCS component has more potential than the Nursing Simulation Laboratory (NSL) component to be problematic in relation to scoring, and a study of the NSL component was more recently conducted (Bosco, 2003). The higher potential for problems with accuracy and consistency in scoring the PCS component exists because of the complexity of administering a high-stakes examination in the authentic patient-care environment. Scored elements that entail the raters evaluating candidates’ recordings of patient data on a recording form were not included as part of the present study so that the raters could focus exclusively on the performance of nursing care by the candidates. Performance of candidates caring for patients on the video
stimulus depicted accurate performance as well as deliberately scripted errors in performance.

Scoring video-recorded stimuli is considered the only feasible option for this study due to patient privacy laws within the authentic patient care environment where the examination is actually administered. In addition, creating a simulated stimulus for scoring allowed the researcher to create situations with the highest degree of comparability and consistency in test items among candidates. Comparability and consistency is assured during actual examinations through the application of examination administration protocols that identify criteria for patient selection (e.g., in terms of the acuity of the patient’s condition) and type of nursing care required as well as the type and number of areas of care assigned to each candidate during each patient care situation.

The data collected are both categorical and continuous. The scoring of each critical element and area of care results in dichotomous data; that is, each critical element and area of care is scored pass/fail and ultimately results in an absolute determination of pass/fail for the examination. Thus, for the purposes of addressing each research question, raters’ determinations for each critical element and area of care is scored as either accurate or inaccurate based on predetermined scoring criteria and candidate performance. Rater characteristics of RPAC and test site hospital affiliation, full time work role, race, gender, and highest degree earned were also considered categorical data. Rater characteristics of time in clinical examiner role, age, and the frequency of working as a rater within the past twelve months were gathered as continuous data.
The CPNE is a complex performance assessment organized into test items called patient care situations that consist of areas of care and critical elements. Critical elements are discrete, observable behaviors that raters score during each examination. Varying numbers of critical elements compose each area of care. Performance of candidates providing nursing care for patients in the video simulation depicted accurate performance as well as deliberately scripted errors in performance. All raters scored the same performance of the same two candidates. Each candidate was assigned some similar and some different areas of care during patient care situations. Similar areas of care are referred to as overriding areas of care. Overriding areas of care are always scored throughout every examination. Differing areas of care are referred to as selected areas of care. Selected areas of care are chosen by the clinical examiners administering the examination (i.e., raters) based on the needs of the patient and the examination administration protocols. Raters scored each candidate’s performance in the same sequence while viewing the performance during a synchronized webcast of the video simulation.

After receiving instructions from the researcher, raters used a paper and pencil form depicting all required critical elements for each area of care performed by each candidate, similar to the form used during actual CPNE administrations. Current scoring criteria for the CPNE, developed collaboratively by the nursing faculty and expert psychometricians of Excelsior College, were applied by the raters. Scoring criteria require that each candidate who passes the CPNE demonstrate competence on each assigned critical element within overriding and selected areas of care with 100% accuracy. Examination scores result in an absolute decision of pass or fail.
Phases.

The research was conducted in two phases. The first phase involved creating the video simulation depicting performance of two candidates, verifying the fidelity of the simulation, assessing visual and audio quality of the webcast video, and obtaining Institutional Review Board approval from Excelsior College and the State University of New York at Albany. This phase included two pilot studies. The first pilot was conducted using a panel of six current and former Excelsior College associate degree program faculty members and consultants who were deemed unbiased and informed experts. The experts reviewed the script and first version of the simulation video and responded to 11 items measuring various aspects of the fidelity of the simulation in representing an actual CPNE. Fidelity of the video simulation in depicting an actual CPNE, including candidate, rater, and patient performance as well as the testing environment, was measured during the first pilot study and results are described in Appendix A.

The second pilot included testing the audio and visual acuity of the webcast video using a panel of five raters who have worked in the role of rater for over one year. These raters evaluated the audio and visual quality of the video simulation in the webcast format. Four of the five participants in the second pilot study had broadband or DSL access to the Internet. After viewing several differently formatted versions of the video via webcast, all participants with broadband access to the Internet reported that the video and audio were in synch and rated both the audio and visual quality as good. The member who had dial-up access reported that the audio and visual were out of synch, thus making both the audio and visual quality poor. Based on this information, it became apparent that in order to maintain the integrity of the research design, participants would need
broadband or DSL access to the Internet in order to participate in the study. During the second and final phase, a webinar format was used to collect the data from raters as they viewed the performance of each candidate online. Ultimately, data collected were analyzed and interpreted in terms of the research questions.

**Instruments.**

Raters used a scoring tool created by the researcher that was abbreviated to reflect the critical elements depicted in the video stimulus, but similar to the scoring tool used during the actual complete examination. The scoring tool, included as Appendix B: Demographic Data and Scoring Tool, includes all relevant discrete, observable candidate behaviors, termed critical elements, for each area of care that were scored while viewing the performance in the video stimulus. In addition, information was added to the scoring tool to collect data about each rater, including Regional Performance Assessment Center affiliation, primary test site hospital, ethnicity, age, full time work role, highest degree earned, number of months working in the role of CPNE rater, and how often the participant worked as a CPNE rater within the past 12 months.

**Participants.**

The population under study were raters who score the CPNE at the test-site hospitals where the examination was being administered and who agreed to participate in the study. Each test site falls under the auspices of one of three Regional Performance Assessment (RPAC) Centers: the Northern Performance Assessment Center, which oversees test sites in New York and Pennsylvania; the Midwestern Performance
Assessment Center, which oversees test sites in Ohio, Wisconsin, and Texas; and the Southern Performance Assessment Center, which oversees test sites in Georgia.

All raters are titled clinical examiners and are employed as adjunct faculty at Excelsior College or consultant faculty at the Southern Performance Assessment Center. All raters must meet criteria for appointment (Yarbrough et al., 2007). Criteria for initial and ongoing appointment to the role of clinical examiner include holding a master’s or doctoral degree in nursing and having recent teaching experience with pre-licensure or newly licensed registered nurses in the medical and surgical nursing content covered by the examination. All raters complete a comprehensive training program to learn the principles and concepts relative to competency-based education, performance assessment, and summative evaluation. In addition, they learn the responsibilities of the clinical examiner (i.e., rater) role, including all of the standardized administration and scoring protocols during a mandatory three-phase orientation process. The raters in this study have varying amounts of experience scoring the CPNE and come from a variety of primary practice and education backgrounds and settings.

All raters who did not participate in the pilot study for this research were invited to participate. Participation in the study was voluntary and raters were invited to participate as part of an annual update session. Consent to participate was obtained by having participants read an electronically delivered consent statement prior to engaging in the scoring session and then indicating consent to participate. The sample of raters was determined by those who actually submitted completed research instruments. A targeted convenience sample of 71 raters was obtained from test sites out of the total population of
289 raters nationwide, with representation from each RPAC, yielding a 25% response rate. Of the 71 surveys returned, 4 were incomplete yielding 67 usable surveys.

The sample of items (areas of care and critical elements) being scored were depicted in a 20-minute digital recording of two candidates. Combined elements for the two candidates totaled 15 areas of care and 105 critical elements scored by raters as they viewed the simulated examination of each candidate performing critical elements for defined areas of care. One candidate was a Caucasian female and the other was a male of Filipino descent.

**Procedures.**

Raters were instructed to apply all current administration and scoring protocols while scoring the simulated CPNE with the exception of the stopping protocol. Because of the nature of the examination, there are times when examination and scoring protocols require the rater to stop a candidate’s performance, remove the candidate from the patient’s bedside, and make a determination of fail for a PCS. For example, if a candidate’s action may put a patient in physical or emotional jeopardy, the PCS is stopped by the rater and a determination of fail is made after consultation with an examination administrator referred to as the clinical administrator (CA). For the purposes of this study, raters were requested to merely note where they would have stopped the candidate’s performance and the rationale, and then continue scoring the candidate’s performance. That is, they continued scoring the examination even if they may have stopped the candidate in an actual testing situation. Waiver of the stopping protocol allowed each critical element to be scored by all raters.
In summary, the research procedure involved multiple steps including: (1) developing the orientation to the research script and creating the video stimulus, (2) developing the scoring tool for use with raters at the time of data gathering, (3) obtaining Internal Review Board approvals from Excelsior College and the State University of New York at Albany, (4) pilot testing the video and scoring tool with Excelsior College faculty members who have expert knowledge around about CPNE administration and scoring protocols to verify fidelity of the video simulation and scoring tool in representing an actual CPNE as well as (5) pilot testing the video for visual and audio quality of the video stimulus. Raters involved with pilot testing the research were eliminated from subsequent data gathering and analysis, and finally (6) scoring and individual data from raters were gathered via multiple synchronous online scoring sessions, followed by data analysis and reporting.

Data Analysis

Accuracy.

The second research question was addressed by calculating percent agreement. After validating accurate scoring determinations via a pilot study that included evaluation by expert faculty during phase one of the study, percent agreement was used to determine the extent to which raters were accurate in scoring candidate performance. Therefore, because accurate scores were known and because clear criteria for scoring and passing the examination exist, accuracy determinations were analyzed by calculating percent agreement.
Reliability.

Putka, Le, McCloy, and Diaz (2008) point out that the nature of the coefficients used to estimate reliability is intimately related to the measurement design used to produce the scores. Several authors warn against the dangers of Pearson $r$ and exclusively refer to intraclass correlation as the appropriate measures of reliability (Bartko, 1976; Putka, Le, McCloy, & Diaz, 2008). Because in this study each rater is scoring the same candidate’s performance independently and because there is a large number of individual data points for scoring (i.e., $n=105$), it is determined that using intraclass correlation (ICC) as a measure of inter-rater reliability exclusively is a sound approach to addressing the first research question (Walter, Eliasziw, & Donner, 1998).

Putka et al. (2008) emphasize that the structure of the measurement design is extremely important in determining how reliability can be estimated. Determining intraclass correlation appropriately estimates interrater reliability. Intraclass correlation determinations involve analysis of variance (ANOVA) to establish estimates of variance due to factors determined by the study design (Downing, 2004). The analysis of variance approach allows for estimation of the interrater reliability for all raters. Therefore intraclass correlation was calculated to determine the extent to which there was interrater agreement among raters in scoring candidate performance during the CPNE. Although intraclass correlation is preferred over percent agreement determinations in regard to interpretability of reliability coefficients, because it accounts for chance occurrence of agreement, it is not as refined as the application of generalizability (G) theory.
Generalizability theory.

Intraclass correlations estimated to address the first research question will be augmented by G-theory to analyze multiple sources of score variance in addressing the third research question. Downing (2004) emphasizes the fact that complex performance assessments require an equally complex reliability model, which is best accomplished using generalizability theory. Ultimately, the goal of the research design and analysis is to generalize from the sample to the population of raters and the universe of examination scores.

Generalizability theory was applied to partition major sources of variance in CPNE scores and estimate dependability of examination scores. In other words, generalizability theory was applied to obtain information about variance components, which can ultimately be used to improve the CPNE. Application of generalizability theory involves determinations of ANOVA, but allows for estimating sources of variance in relation to how consistently findings from a sample of cases can be generalized to a universe of cases (Downing, 2004). The goal of any generalizability study is to identify as many variance components as possible (Brennan, 2001; Brennan & Kane, 1979; Shavelson & Webb, 1991).

The purpose of applying generalizability theory in this study was to quantify, via the calculation of G-coefficients and standard errors of measurement (SEM), the sources of variance for CPNE scores. Some of the variance in CPNE scores will be due to various facets incorporated in the G-study design, such as characteristics of raters and items. The G-theory model used in the present research complements the model used in an analysis of the CPNE conducted in 2005 that revealed a high degree of consistency over time in
examination outcomes as well as acceptable levels of dependability (Yarbrough et al., 2007).

Generalizability theory is useful in this type of analysis because typical measures of reliability for the CPNE would be somewhat unsatisfactory; they would not take into account the variance of the individual components of the scores (i.e., rater, candidates, and items). Although candidate variance is expected, and it is in fact what we want to measure, the variance due to raters or items is considered a type of error variance. That is, there would be less confidence in the CPNE scores if more error variance were reported as part of a candidate’s score.

The current research involved application of a measurement design focused on ratings of candidate performance as judged by raters as the object of measurement. That is, candidates are the objective of measurement. The facets of raters, candidates, and items were analyzed using both crossed and nested designs. It is important to note that all candidates were scored by all raters and all raters scored all items. Some items were the same for each candidate (i.e., overriding area of care critical elements) and some items differed between candidates (i.e., selected area of care critical elements). In cases where raters scored similar items for each candidate, items were crossed with raters (i.e., for items within overriding areas of care). In cases where raters were scoring different items for each candidate, items were nested within candidates and crossed with raters (i.e., for items within selected areas of care). This type of model decomposes sources of variability in CPNE scores with a design that in some cases is depicted by raters crossed with items nested within candidates (i.e., r x i:c). The sources of variability using this type of a model include: (1) candidates (c), (2) raters (r), and (3) items (i), and (4) residual random
effects (i.e., sources of error variance of an unknown origin). All components are samples of conditions and therefore facets within the design are treated as “random” conditions drawn from an imaginary universe of conditions (Shavelson & Webb, 1991).

Having confidence in decisions made based on examination scores is contingent on having sufficient evidence to support generalization from a sample score to a universe score. Shavelson and Web (1991) point out that while test users really want to know each person’s universe score, the best that can be done is to determine the accuracy of desired generalizations from the test scores. The outcome of this research will add to the existing body of validity evidence for inferences made from CPNE scores and will address testing technical standards related to accurate score reporting and interpretation (AERA, APA, & NCME, 1999; Becker & Pomplun, 2006).

Sample size and power.

Considering sample size in relation to power, a priori power analysis for a desired power level of .9 at a .05 significance level when scoring 105 items revealed a required sample size estimation of 11 raters, well within what was expected for the present study and ultimately achieved. For this study, with 67 raters scoring 105 items, sample size is more than adequate and sufficient power, which is the ability to detect a significant difference if one is present, was acceptable for statistical analysis (Thorndike, 2005). Specific to calculating intraclass correlation (ICC), optimum samples size is a function of the size of the ICC and the number of ratings per subject as well as the desired significance level and the desired width of the confidence interval. With a desired confidence level of .95, a desired power level of .80, and 20 ratings, a sample size of 2 to 11 is required (Bartko, 1976), well within the specifications of the present study.
In addition, the sample size in this study is adequate for the purpose of applying generalizability theory. When applying generalizability theory, the power issue is considered from the perspective of standard errors for the estimated variance components. That is, if the standard error of the variance component is large relative to the magnitude of the estimated variance component (i.e., for raters, candidates and items) then there is reason to wonder about whether the sample size is large enough for meaningful decisions (Brennan, 2001). In the present study, standard errors for all points of data analysis were acceptable.
Chapter Four: Results

Data obtained from trained raters were analyzed in relation to the representativeness of the sample and each of the three research questions driving the study: (1) To what extent is there interrater agreement among raters in scoring candidate performance during the CPNE?; (2) To what extent are raters accurate in scoring candidate performance during the CPNE?; (3) What are the major sources of variance in CPNE scores?

Identifying the extent to which there is interrater agreement among raters in scoring candidate performance during the CPNE was examined by estimating the intraclass correlation. Also, because the accurate or “true” answer to each item was known, it was possible to calculate a percent accuracy score for each item to determine to what extent raters were accurate in scoring candidate performance during the CPNE. Correct performance by the candidate was determined by the rater if the candidate’s behavior met the pre-determined criteria for passing each critical element in assigned selected areas of care as well as all overriding areas of care. Critical elements are discrete, observable behaviors and thus are the focus of scoring for CPNE, which is a criterion-referenced performance examination. Accurate scoring by the rater was determined if the rater scored the candidate’s performance in accordance with scoring criteria and made a scoring determination that was correct. Finally, major sources of variance in CPNE scores were estimated by applying generalizability theory and estimating G-coefficients and standard errors of the variance components and G-coefficients.
Characteristics and Representativeness of the Sample

Descriptive analysis was done to identify the characteristics of the sample and to evaluate the sample in relation to the population of CPNE raters on specified dimensions. Tables 1-3 illustrate that the sample was determined to be comparable to the population of CPNE raters in terms of gender and Regional Performance Assessment Center (RPAC) affiliation. A comparison of number of months in the rater role reveals that the sample had a disproportionate number of newer raters, but a proportionate number of raters in the role for 50 months or more.

Table 1: Gender

<table>
<thead>
<tr>
<th></th>
<th>% Sample</th>
<th>% Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>98.5</td>
<td>96.5</td>
</tr>
<tr>
<td>Male</td>
<td>1.5</td>
<td>3.5</td>
</tr>
</tbody>
</table>

Table 2: RPAC Affiliation

<table>
<thead>
<tr>
<th>RPAC</th>
<th>% Sample</th>
<th>% Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Midwestern</td>
<td>55.2</td>
<td>55.4</td>
</tr>
<tr>
<td>Northern</td>
<td>35.8</td>
<td>33.1</td>
</tr>
<tr>
<td>Southern</td>
<td>9.0</td>
<td>13.5</td>
</tr>
<tr>
<td>Total</td>
<td>100.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>
Table 3: Number of Months in the Rater Role

<table>
<thead>
<tr>
<th># Months</th>
<th>% Sample</th>
<th>% Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 to 24</td>
<td>41.8</td>
<td>23.7</td>
</tr>
<tr>
<td>25 to 49</td>
<td>13.4</td>
<td>28.5</td>
</tr>
<tr>
<td>50 to 73</td>
<td>1.5</td>
<td>14.5</td>
</tr>
<tr>
<td>&gt;73</td>
<td>43.3</td>
<td>32.8</td>
</tr>
</tbody>
</table>

Tables 4-6 include data on the race/ethnic, full-time work role, and number of times raters worked in the role of clinical examiner within the past year. No similar data for the population were available for comparison to the sample.

Table 4. Race/Ethnicity of Raters in the Sample

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>62</td>
<td>92.5</td>
</tr>
<tr>
<td>Black</td>
<td>4</td>
<td>6.0</td>
</tr>
<tr>
<td>Asian</td>
<td>1</td>
<td>1.5</td>
</tr>
<tr>
<td>Total</td>
<td>67</td>
<td>100.0</td>
</tr>
</tbody>
</table>
### Table 5: Full-time Work Role of Raters in the Sample

<table>
<thead>
<tr>
<th>Role</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Educator</td>
<td>37</td>
<td>55.0</td>
</tr>
<tr>
<td>Staff Development &amp; “Other”</td>
<td>6</td>
<td>11.0</td>
</tr>
<tr>
<td>Administrator</td>
<td>9</td>
<td>13.0</td>
</tr>
<tr>
<td>Advanced Practice Nurse</td>
<td>10</td>
<td>15.0</td>
</tr>
<tr>
<td>Retired</td>
<td>5</td>
<td>6.0</td>
</tr>
<tr>
<td>Total</td>
<td>67</td>
<td>100.0</td>
</tr>
</tbody>
</table>

### Table 6. Number of Times Raters in the Sample Worked in the Past 12 Months

<table>
<thead>
<tr>
<th>Times Worked</th>
<th>Frequency</th>
<th>Percent</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤ 20</td>
<td>52</td>
<td>77.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt; 20</td>
<td>15</td>
<td>22.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>67</td>
<td>100.0</td>
<td>3</td>
<td>40</td>
<td>14.6866</td>
<td>8.3052</td>
</tr>
</tbody>
</table>

In summary, the sample is comparable to the population in relation to gender and affiliations with Regional Performance Assessment Centers. The majority of the raters in the sample are white; no Hispanic raters were represented. There is wide variability in the amount of time the raters in the sample have worked in the role of clinical examiner. Some have worked in the role as little as two months and others are extremely seasoned in the role after more than 28 years. There were large differences in the number of times raters in the sample worked in the clinical examiner role within the past year. The
majority of raters in the sample worked 20 times or less in the role of clinical examiner within the past year, while less than a quarter of the sample worked more than 20 times within the past year.

**Research Question 1: To what extent is there interrater agreement among raters in scoring candidate performance during the CPNE?**

The first research question examines the extent to which there is inter-rater agreement among raters in scoring candidate performance during the CPNE. The extent to which there is inter-rater agreement among raters in scoring the performance of both candidates on all items was estimated by determining the intraclass correlation (ICC). Data analysis revealed a significant intraclass correlation of .952, with a 95% confidence interval (lower bound .938, upper bound .964, width .026). An ICC of .952 indicates high scoring consistency or homogeneity among raters.

**Research Questions 2: To what extent are raters accurate in scoring candidate performance during the CPNE?**

The second research question investigates the extent to which raters are accurate in scoring candidate performance during the CPNE. Data addressing this question were analyzed on a number of levels, including the overall accuracy of raters in determining each candidate’s examination score, accuracy in scoring critical elements within various areas of care, and accuracy in terms of the nature of the scoring error (i.e., inaccurately scoring failed elements as “pass” and inaccurately scoring passed elements as “fail”).

The most problematic scored element and the outlier in terms of scoring accuracy was the critical element that requires the candidate to assess equality of pupil size. Only 37% of raters scored this critical element accurately. This error was not a preplanned and
scripted error, but rather was an artifact of the videotaping. That is, during the editing process, the candidate’s accurate performance was only partially included on the captured video, resulting in incomplete and thus failing performance. The unplanned error in performance was discussed with faculty experts around the CPNE who viewed the video and determined that the candidate’s performance was clearly depicted in the video and clearly did not meet the criteria for passing. Therefore, percent accuracy data and ANCOVA analysis were analyzed for candidate two both with and without the outlier. Data analysis including the outlier is depicted in Appendix C.

Overall accuracy reflects the broadest view of rater precision. This measure was addressed by evaluating the percent of raters who scored each candidate’s overall performance accurately. That is, was the overall determination by the rater for the Patient Care Situation (PCS) 100% accurate? Figure 2 illustrates that 79.1% of raters scored candidate one accurately and Figure 3 illustrates that 68.7% of raters scored student two accurately. Of all raters, 91.0% scored at least one of the two candidates’ performance accurately, and 56.7% scored both candidates’ performance accurately.

Figure 2: Percent Accuracy of Rater Scoring: Candidate One
Figure 3: Percent Accuracy of Rater Scoring: Candidate Two

Accuracy in scoring examination critical elements within overriding and selected areas of care is a more fine-grain approach to considering rater accuracy. Table 7 depicts percent accuracy for all overriding and selected area of care scored critical elements.

Table 7: Percent of Critical Elements Accurately Scored

<table>
<thead>
<tr>
<th></th>
<th>Within Overriding Areas of Care</th>
<th>Within Selected Areas of Care</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>60.00%</td>
<td>73.58%</td>
</tr>
</tbody>
</table>

Finally, percent accuracy and nature of the scoring error for specific critical elements is another way to view accuracy data. Table 8 identifies percent accuracy of scored elements that were scored with 97% accuracy or less, including the nature of the inaccuracies in scoring.

Seven of the ten elements most commonly scored inaccurately involved raters scoring passed items as failed. In terms of the intended uses of CPNE scores, this is the least detrimental type of error in terms of patient safety and upholding the stringent standards of the nursing profession. However, from a candidate’s perspective on high-stakes testing, this type of error is problematic.
Table 8: Percent Accuracy for Scored Elements and Nature of Scoring Errors

<table>
<thead>
<tr>
<th>Area of Care</th>
<th>Critical Element</th>
<th>Percent Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Candidate 1 and 2</strong></td>
<td><strong>2.</strong> Protects self, others, and the environment from contamination</td>
<td>93</td>
</tr>
<tr>
<td>Asepsis</td>
<td><strong>4.</strong> Disposes of contaminated material in the designated container</td>
<td>97</td>
</tr>
<tr>
<td>Physical Jeopardy</td>
<td><strong>Any action or inaction on the part of the students that threatens patient’s physical well-being.</strong></td>
<td>97</td>
</tr>
<tr>
<td><strong>Candidate 1</strong></td>
<td><strong>4.b.</strong> Assesses patient’s respiratory status before initiating respiratory hygiene activities by auscultating breath sounds over upper and lower lobes by systematically moving the stethoscope from side to side</td>
<td>94</td>
</tr>
<tr>
<td>Respiratory Management</td>
<td><em>Reassesses respiratory status immediately after respiratory hygiene activities</em></td>
<td>93</td>
</tr>
<tr>
<td>Medications</td>
<td><strong>9.c.</strong> Clears air from the tubing before initiating flow</td>
<td>97</td>
</tr>
<tr>
<td><strong>Candidate 2</strong></td>
<td><strong>2.a.</strong> Compares the extremities by palpating for the presence or absence of most distal pulses</td>
<td>84</td>
</tr>
<tr>
<td>Peripheral Vascular Assessment</td>
<td><strong>1.</strong> Complies with established guidelines</td>
<td>88</td>
</tr>
<tr>
<td>Neurological Assessment</td>
<td><strong>4.a./b.</strong> Assesses pupillary response regarding equality of pupil size and reaction to light</td>
<td>37</td>
</tr>
<tr>
<td>Abdominal assessment</td>
<td><strong>5.</strong> Performs light palpation over all 4 quadrants for tenderness or rigidity, unless contraindicated</td>
<td>97</td>
</tr>
</tbody>
</table>

* raters scored failing performance as a “pass”.

** raters scored passing performance as a “fail”.

¹ data analysis with outlier depicted in Appendix C.
Because the candidate’s performance was clearly depicted in the video and clearly did not meet the criteria for passing, it is interesting to speculate why this element was so problematic. Two possible reasons for the high percentage of inaccuracy are that (1) having a candidate only assess one pupil instead of both pupils is not a common error, and (2) if the raters had not clearly picked up the error in performance during an actual CPNE administration, they may have relied on the candidate’s recording of the assessment data to verify what they thought they saw. In the case of this study, recording of assessment data was not included in the design and thus, the raters could not refer to it.

Finally, rater cohort accuracy in relation to RPAC affiliation is another way to view the data. Table 9 depicts rater accuracy for each candidate per RPAC.

Table 9: Accuracy of Raters per RPAC

<table>
<thead>
<tr>
<th>RPAC</th>
<th>Number of Raters</th>
<th>% Raters Scoring Candidate One Accurately</th>
<th>% Raters Scoring Candidate Two Accurately</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPAC</td>
<td>37</td>
<td>81%</td>
<td>68%</td>
</tr>
<tr>
<td>NPAC</td>
<td>24</td>
<td>71%</td>
<td>71%</td>
</tr>
<tr>
<td>SPAC</td>
<td>6</td>
<td>100%</td>
<td>67%</td>
</tr>
</tbody>
</table>

Chi square analyses were performed to test the hypothesis that scoring accuracy differed between RPACs. The $\chi^2$ statistics did not reach a level of significance at the $p < .05$ level for the scoring of either candidate (candidate 1: $\chi^2 = 2.666; \text{df} = 2$; candidate 2: $\chi^2 = .084; \text{df} = 2$). This analysis reveals that there was no dependence between scoring accuracy and the raters’ RPAC affiliation.
Research Question 3: What are the major sources of variance in CPNE scores?

Another interesting approach to the data is via analysis of covariance (ANCOVA). Preliminarily analyzing the data in this way prior to generalizability analysis helps to discern what facets should be included in the analysis. Specifically, it tests the hypothesis that raters who work more frequently in the role of rater are more accurate in scoring the CPNE and that raters with more months in the role of rater are more accurate. This type of analysis controls for variance in one variable, while analyzing variance on another variable. Results from the ANCOVA are presented in Table 10 and Table 11, which illustrate data analysis for raters’ accuracy in scoring each candidate analyzed with number of times worked in the past year and length of time in the role of rater as covariates.
Table 10: ANCOVA for Rater Accuracy in Scoring Candidate One

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
<th>Partial Eta Squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corrected Model</td>
<td>7.750</td>
<td>45</td>
<td>.172</td>
<td>1.088</td>
<td>.430</td>
<td>.700</td>
</tr>
<tr>
<td>Intercept</td>
<td>6.890</td>
<td>1</td>
<td>6.890</td>
<td>43.527</td>
<td>.430</td>
<td>.675</td>
</tr>
<tr>
<td>Number of times worked in the past year</td>
<td>1.726</td>
<td>1</td>
<td>1.726</td>
<td>10.901</td>
<td>.003</td>
<td>.342</td>
</tr>
<tr>
<td>Number of months in the role of rater</td>
<td>7.490</td>
<td>44</td>
<td>.170</td>
<td>1.075</td>
<td>.442</td>
<td>.693</td>
</tr>
<tr>
<td>Error</td>
<td>3.324</td>
<td>21</td>
<td>.158</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>53.000</td>
<td>67</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>11.075</td>
<td>66</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R Squared = .700 (Adjusted R Squared = .057)

Analysis of covariance in rater scoring of candidate one with times worked and number of months in the rater role analyzed as covariates revealed that the covariate, (number of months in the rater role), was not significantly related to the rater accuracy in scoring, $F(44, 21) = 1.075$, $p<.05$. However, the other covariate (number of times worked in the past year), was significantly related to the rater accuracy in scoring, $F(1,21)= 10.901$, $p<.05$ with a moderate effect size.
Table 11: ANCOVA for Rater Accuracy in Scoring Candidate Two

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
<th>Partial Eta Squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corrected Model</td>
<td>51.230</td>
<td>45</td>
<td>1.138</td>
<td>.647</td>
<td>.890</td>
<td>.581</td>
</tr>
<tr>
<td>Intercept</td>
<td>5843.963</td>
<td>1</td>
<td>5843.963</td>
<td>3321.449</td>
<td>.000</td>
<td>.994</td>
</tr>
<tr>
<td>Number of times worked in the past year</td>
<td>.051</td>
<td>1</td>
<td>.051</td>
<td>.029</td>
<td>.866</td>
<td>.001</td>
</tr>
<tr>
<td>Number of months in the role of rater</td>
<td>48.620</td>
<td>44</td>
<td>1.105</td>
<td>.628</td>
<td>.904</td>
<td>.568</td>
</tr>
<tr>
<td>Error</td>
<td>36.949</td>
<td>21</td>
<td>1.759</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>137207.000</td>
<td>67</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>88.179</td>
<td>66</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$R^2 = .581$ (Adjusted $R^2 = -.317$)

Analysis of covariance in rater scoring of candidate two with times worked and months in role analyzed as covariates revealed that neither covariate was significantly related to the rater accuracy in scoring $F(1, 21) = .704$, $p<.05$ and $F(44, 21) = 1.564$, $p<.05$.

While a level of significance was reached for rater accuracy for candidate one and number of times worked in the past year, it was inconsistent with findings in past research and not supported by the findings for rater accuracy in scoring candidate two, in analyses both with and without the outlier (Nettleton, 1997). For these reasons, neither of the factors analyzed via ANCOVA were included in subsequent analysis.
The final research question, which calls for identification of the major sources of variance in CPNE scores, lends itself to the application of generalizability theory and analysis of G-coefficients and standard errors. G theory expresses reliability coefficients as functions of variance components. Variance components represent the contributions of different sources of variance to the ratings of candidates; the reliability coefficient indicates the proportion of observed score variance that is due to true differences between the candidates attributable to a particular facet. In the present study, raters, candidates, and items nested with candidates were the sources of variances investigated with candidates conceptualized as the object of measurement. The variance components for these facets were estimated using the GENOVA software program (Brennan, 2001). GENOVA is a FORTRAN computer program for univariate generalizability analysis with capabilities for both generalizability (G) and decision (D) studies.

In a design that involved raters (r) crossed with (Area of Care) items (i) and items-nested-within-candidates (i:c), \([r \times (i:c)]\), five sources of variance were evaluated. Percent of variance attributable to the true measureable differences between raters is indicated by “r” and represents variance not otherwise explained by other sources of error. Candidate variance indicated by “c” is the degree to which candidates themselves are the source of variance.

Results from this analysis are presented in Table 12. The main effect of raters (r) accounted for 2.3% of the variance, indicating that on average differences between the raters were small. The items-nested-within-candidate facet (i:c), which is expected to vary among themselves, accounted for an appreciable source of error at 20.7% of the total variance. The candidate variance in this model represents a relatively small source of
error at 1.5%. Residual random effects (i.e., ri:c) claim the largest source of error at 77.8%.

Table 12: Generalizability Study Results for All Area of Care Items

<table>
<thead>
<tr>
<th>Effect</th>
<th>df</th>
<th>T</th>
<th>SS</th>
<th>MS</th>
<th>VC</th>
<th>% of Variance</th>
<th>SEVC</th>
</tr>
</thead>
<tbody>
<tr>
<td>r</td>
<td>66</td>
<td>1169.867</td>
<td>4.9672</td>
<td>0.0753</td>
<td>0.00164</td>
<td>2.3</td>
<td>.0405</td>
</tr>
<tr>
<td>c</td>
<td>1</td>
<td>1166.488</td>
<td>1.5885</td>
<td>1.5885</td>
<td>0.0011</td>
<td>1.5</td>
<td>.0332</td>
</tr>
<tr>
<td>i:c</td>
<td>13</td>
<td>1180.059</td>
<td>13.5717</td>
<td>1.0439</td>
<td>0.01475</td>
<td>20.7</td>
<td>.1214</td>
</tr>
<tr>
<td>rc</td>
<td>66</td>
<td>1174.804</td>
<td>3.3484</td>
<td>0.0507</td>
<td>*0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ri:c</td>
<td>858</td>
<td>1236.000</td>
<td>47.6247</td>
<td>0.0555</td>
<td>0.05551</td>
<td>77.8</td>
<td>.2356</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>0.07136</td>
<td>100.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*This value was a negative variance component which was rounded to zero (Brennan, 2001)

r-raters, c=candidates, i=items; T=total sum of squares, SS=deviation sum of squares, MS= mean square, VC=estimated variance component, SEVC=standard error of variance component computed as the square root of the variance component.

Area of Care level score and standard error calculations:

G coefficient of .89

\[
0.00164 + 0.05551/(2 \times 15) = 0.05715/30 = 0.001905
\]

\[
0.01475/(0.01475 + 0.001905) = 0.01475/0.016655 = 0.88562
\]

Standard error of the G-coefficient of .001904

\[
[(1 - 0.88562)/0.88562) \times (0.01474)] = 0.1291524 \times 0.01474 = 0.001904
\]

Confidence interval

\[
0.001904 \times 1.96 - .89 = 0.886268 \text{ (lower bound)}
\]

\[
0.001904 \times 1.96 + .89 = 0.893732 \text{ (upper bound)}
\]

The G coefficient for CPNE scores (i.e., area of care scores) of .89 with a standard error of .0019 with a 95% confidence interval (lower bound .886, upper bound .894, width .008) indicate that the CPNE measures of Area of Care items demonstrated
acceptable psychometric characteristics. G coefficients of 0.7 or higher are considered acceptable (Brennan, 2001). These results suggest that CPNE scores at the area of care level are dependable estimates.

In a design that involved raters (r) crossed with (critical element) items (i) and items-nested-within-candidates [r x (i:c)], five sources of variance were evaluated. Percent of variance attributable to the true measureable differences between raters is indicated by “r” and represents variance not otherwise explained by other sources of error. Candidate variance indicated by “c” is the degree to which candidates themselves are the source of variance.

Results from this analysis are presented in Table 13. Again in this model, both the rater and the candidate variances represent a relatively small source of error. The main effect of raters (r) accounted for .2% of the variance, indicating that on average differences attributable to the raters were small. The main effects for candidates accounted for .6% of the variance, indicating that on average differences in scores related to the candidate was small. Items nested within candidates (i.e., i:c) represents an appreciable source of error at 27.9%. Residual random effects (i.e., ri:c) claim the largest source of error (71%).
Table 13: Generalizability Study Results for All Critical Elements

<table>
<thead>
<tr>
<th>Effect</th>
<th>df</th>
<th>T</th>
<th>SS</th>
<th>MS</th>
<th>VC</th>
<th>% of Variance</th>
<th>SEVC</th>
</tr>
</thead>
<tbody>
<tr>
<td>r</td>
<td>66</td>
<td>6232.6354</td>
<td>1.0494</td>
<td>0.0159</td>
<td>0.00003</td>
<td>.2</td>
<td>.0055</td>
</tr>
<tr>
<td>c</td>
<td>1</td>
<td>6232.1765</td>
<td>0.5905</td>
<td>1.5905</td>
<td>0.00009</td>
<td>.6</td>
<td>.0095</td>
</tr>
<tr>
<td>i:c</td>
<td>94</td>
<td>6260.5821</td>
<td>28.4056</td>
<td>0.3022</td>
<td>0.00435</td>
<td>27.9</td>
<td>.0659</td>
</tr>
<tr>
<td>rc</td>
<td>66</td>
<td>6234.1101</td>
<td>0.8841</td>
<td>0.0134</td>
<td>0.00005</td>
<td>.3</td>
<td>.0071</td>
</tr>
<tr>
<td>ri:c</td>
<td>6204</td>
<td>6331.0000</td>
<td>68.4843</td>
<td>0.0110</td>
<td>0.01104</td>
<td>71.0</td>
<td>.1051</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>0.01556</td>
<td></td>
<td></td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

r=raters, c=candidates, i=items; T=total sum of squares, SS=deviation sum of squares, MS= mean square, VC=estimated variance component, SEVC=standard error of variance component computed as the square root of the variance component.

Area of Care level score and standard error calculations:

G coefficient of .99

0.00003 + 0.01104/(2 x 96)= 0.01107/192 = .0000576
0.00435/(0.00435 + .0000576) = 0.00435/.0044044 = .9876487

Standard error of the G coefficient of .000054

[(1 - .987648)/.987648) x (0.00435)] = 0.0125064 x 0.00435 = .000054

Confidence Interval

.000054 x 1.96 - .99 = .9898942 (lower bound)
.000054 x 1.96 + .99 = .9901058 (upper bound)

The G coefficient for CPNE scores (i.e., all critical elements) of .99 with a standard error of .000054 with a 95% confidence interval (lower bound .9898, upper bound .9901, width .0003) indicates that the CPNE measures of critical elements within all areas of care demonstrated acceptable psychometric characteristics. G coefficients of 0.7 or higher are considered acceptable (Brennan, 2001). These results suggest that CPNE scores at the critical element level are dependable estimates.
The G coefficients, standard errors, and confidence intervals for the variance components for the facets measured in this study are depicted in Table 14.

Table 14: G-Coefficients, Standard Error of G-coefficients, and Confidence Intervals

<table>
<thead>
<tr>
<th>CPNE Measure</th>
<th>G-Coefficient</th>
<th>Standard Error</th>
<th>95 % Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area of Care</td>
<td>.89</td>
<td>.00190</td>
<td>lower bound .886, upper bound .894, width .008</td>
</tr>
<tr>
<td>Critical Element</td>
<td>.99</td>
<td>.00005</td>
<td>lower bound .989, upper bound .990, width .0003</td>
</tr>
</tbody>
</table>

G-coefficients are high and standard errors are low relative to the variance components. The width of the 95% confidence intervals reveals high precision suggesting that the variance components are stable estimates.
Chapter Five: Conclusion and Implications

Summary of Findings

It is important to continually add to the body of information regarding validity evidence and threats for the CPNE and substantiate the use of the examination scores for their intended purpose. Toward this end, the researcher established measures of consistency and accuracy of the raters scoring video-based, simulated candidate performance. This was accomplished via a method that entailed 67 raters scoring the performance of two candidates during a video-taped simulated CPNE. In addition, the study involved determining the dependability of scores by partitioning and estimating the major sources of variance in scores.

Raters in this study were representative of the entire population of raters across the United States who administer and score the CPNE in terms of gender and affiliations with RPAC. There was wide variability in the amount of time raters in both the sample and population have worked in the role of clinical examiner (i.e., rater). A comparison of the number of months in the rater role revealed that the sample had a disproportionate number of newer raters when compared to the population. Analysis of covariance was done to test the hypothesis that raters with more months in the role of rater are more accurate in scoring. The hypothesis was not supported by prior research (Nettleton, 1997) and was not supported in the ANCOVA for candidates one or two in this study. For these reasons, this facet was not included in subsequent generalizability analysis. While it may seem intuitive that more months in the role of rater will result in increased scoring accuracy, this hypothesis was also not supported by anecdotal evidence obtained from experts who observed and provided feedback on the CPNE in August and September,
2010. These experts commented that newer raters appeared to adhere more closely to the scoring protocols than more seasoned raters, and thus may be more likely to score with increased accuracy.

To what extent is there interrater agreement among raters in scoring candidate performance during the CPNE? The extent of interrater agreement in scoring candidate performance was estimated using intraclass correlation (ICC), which revealed high scoring consistency or homogeneity among raters (i.e., intraclass correlation of .952, with a 95% confidence interval, lower bound .938, upper bound .964, width .026). The fact that the CPNE was scored by raters with a high degree of consistency is extremely important because the CPNE is a high-stakes examination administered at 20 test site hospitals located throughout the United States by 289 clinical faculty members and consulting faculty who serve in the role of rater. In terms of the validity of the examination scores, it is very important that each candidate’s performance is scored consistently by raters regardless of where the candidate takes the examination. In other words, the raters within the sample can be viewed as interchangeable.

To what extent are raters accurate in scoring candidate performance during the CPNE? Of the 105 critical elements scored by raters in this study, 95 were scored with 98%, 99%, or 100% accuracy. Only ten critical elements were scored with 97% accuracy or lower. Scoring accuracy of critical elements ranged from 37% to 100%. With the outlier removed, the lower end of the accuracy range was 84%. It is important to note that the critical element that resulted in 37% scoring accuracy was an outlier in terms of scoring accuracy and was not originally planned or scripted as part of the design. Because the error was not in the planned design, yet was clearly depicted in the video and did not
meet the scoring criteria for passing, data analysis was conducted both with and without the outlier for candidate two. An important finding of this study is that 91% of critical elements were scored with a high degree of accuracy (i.e., 97% or higher scoring accuracy). This level of accuracy is acceptable for a performance assessment that requires the candidate to provide nursing care to actual patients in the authentic patient care environment.

At the level of the PCS, the extent of rater accuracy in scoring candidate performance during the CPNE varied somewhat between candidate one and candidate two (i.e., candidate 1 was scored accurately by 79.1% of raters and candidate 2 was scored accurately by 68.7% of raters). Even though rater accuracy at the critical element level was acceptable and the vast majority of scoring errors that did occur involved scoring passed critical elements as failed, which are less detrimental errors in scoring from a patient safety perspective, a higher rate of scoring accuracy is something that can and should be addressed through enhanced rater training.

It is important to consider why the raters scored the performance of candidate 2 less accurately. It may be that the errors scripted into the second candidate’s performance were more difficult to identify for a variety of reasons, despite findings from the pilot study to the contrary. One possibility is that rater fatigue may have played a role in decreased scoring accuracy from candidate one to candidate two because all the raters scored the second candidate after the first candidate. In this study, 91.0% of raters scored one of the two PCS’s accurately. It is important to note that a margin for rater error is built into the actual examination, which allows the student to repeat one adult and one child PCS.
The nature of the scoring errors is another important point of consideration. Of the ten critical elements scored with 97% or less accuracy, seven were errors that involved raters scoring passing performance as a fail. While this is a less detrimental type of error in terms of assuring the rigorous standards of the profession, is quite detrimental to candidates taking such a high-stakes and costly examination. It is important to note that CPNE administration protocols require a clinical associate (CA) to act in the role of examination administrator throughout each examination. It is the CA’s responsibility to consult with the rater and verify the point of failure prior to the rater officially determining the failure during a PCS. Substantiation of the failure by the CA includes verifying that the rater accurately adhered to the examination administration and scoring protocols before the rater can officially deem that the candidate has failed the PCS. One of the core functions of the CA is to validate and process failures before any final determination is made on the outcome of a PCS. Failures are not final until the CA collaborates with the rater, speaks to the candidate and then confirms the decision. In cases where a rater inaccurately scored passing performance as failing, the CA would intervene and either have the rater rescoring the critical element or grant the candidate a repeat PCS without penalty. Given the high-stakes nature of the examination, this type of validation of the scoring determinations by raters is necessary and appropriate to ensure fair testing conditions (AERA, APA, NCME, 1999).

Out of a total of 105 scored critical elements, three errors in scoring critical elements involved raters scoring failing performance as passing (i.e., 2.9%). This is the most detrimental type of error in terms of substantiating the interpretive argument for examination scores. The outlying critical element in terms of scoring accuracy, which
required the candidate to assess pupillary response, was one that a large number of raters scored as a pass when the candidate actually failed the critical element. Other critical elements with low accuracy scores and a scoring error of this nature (i.e., scoring failing performance as a pass) were ones that require the candidate to reassess respiratory status immediately after respiratory hygiene activities and to compare the extremities by palpating for the presence or absence of the most distal pulses. These types of scoring errors need to be minimized to the extent possible and is best achieved by decreasing ambiguity in the critical element and scoring criteria as well as enhancing rater training. In the case of the particular critical elements referenced here, ambiguity in the critical elements and scoring criteria is not apparent and thus the most crucial intervention is rater training.

Chi square analysis revealed that there was no dependence between scoring accuracy and RPAC affiliation of the raters. This is an important finding because it reveals that there does not appear to be any factor influencing accuracy in scoring attributable to RPAC affiliation. In addition, it is helpful to know that it will be efficient and effective to distribute resources and interventions aimed at enhancing rater accuracy equally across all RPACs.

*What are the major sources of variance in CPNE scores?* G-theory was successfully applied to provide information about variance components. G-theory was used because typical measures of reliability for the CPNE are somewhat unsatisfactory as they do not take into account the variance of the individual components of the examination.
In the G-study designs employed, five sources of variance were evaluated at the area of care and critical element levels. Analysis of scoring for both candidates revealed that main effects for raters and candidates were small, with appreciable sources of error attributable to items-nested-within-candidates and residual random effects. The main effect of raters accounted for 2.3% and .2% of the variance respectively. The main effect of candidates accounted for 1.5% and .6% respectively. The item-nested-within-candidate facet accounted for an appreciable amount of error variance 20.7% and 27.9% respectively. Residual random effects claimed the largest source of error at 77.8% and 71% respectively. Although some candidate and rater variance is expected, there would be less confidence in CPNE scores for their intended purpose if more variance were reported as part of rater or candidate facets. Overall, the G-study results suggest that CPNE scores at the area of care and critical element levels are dependable estimates (i.e., .89 with a standard error of .0019, and .99 with a standard error of .000054, both at the 95% level of confidence). These findings are a substantial contribution to the body of validity evidence supporting the interpretive argument for the use of CPNE scores because it is important to understand fully the nature of the major facets factoring into scores.

**Limitations and Delimitations**

This study has a number of limitations and delimitations. It is important to remember that the 67 participants in this study were similar in some ways to the entire population of 289 raters in terms of selected characteristics, but there were more raters with less experience in the role than in the total population of raters, which may have influenced results.
Although the ANCOVA analysis substantiated findings from prior research (Nettleton, 1996) and revealed no significant relationship between length of time in role and accuracy in scoring, this is still an important point of consideration that may require additional analysis. No comparison data between the sample and the population were available in terms of number of times worked in the past twelve months. In the present study, the ANCOVA analysis for this rater characteristic (i.e., number of times worked in past twelve months) revealed conflicting findings between candidate one and candidate two as well as with prior research that also revealed no significant relationship between number of times worked and scoring accuracy (Nettleton, 1996). In the present study, number of times worked in the past twelve months was significantly related to rater accuracy in scoring for candidate one \((F(1,21)=10.901, p<.05\) with a moderate effect size), but was not significantly related to rater accuracy for candidate two for this sample of raters \((F(44,21)=.029, p<.05)\). These conflicting findings both within the study and with prior research require further analysis to determine if number of times worked in the past twelve months is significantly related to rater accuracy. If it there is consistent evidence that it is, it should be included as a facet in future G-studies.

Another limitation results from the fact that the raters volunteered to participate and were not chosen randomly from the population. Because a random sampling plan was not employed, findings cannot be generalized to the entire population of raters because of self-selection bias that potentially occurred in a manner unknown to the researcher.

It is speculated that the largest potential source of error in the design is the fact that simulation was used. Despite high ratings for the fidelity of the video simulation of
the examination, the actual experience of viewing candidate performance on a computer
monitor is much different than observing a candidate’s performance live. Another
potential source of error is the fact that the scoring tool used in the study was somewhat
different than the scoring tool used during the actual examination. The scoring tool used
in the study was abbreviated to only include critical elements that would be scored and
was evaluated as part of pilot #1 (Appendix A). This modification to the scoring tool may
have influenced scoring to a small degree, but is not considered to be a substantial factor
impacting raters’ scoring.

A substantial delimitation of the study is that the validity evidence obtained can
only be applied to the performance portion of the PCS component of the examination.
This is because the nursing simulation component, nursing care planning documentation,
and the candidates’ recordings of patient related data, all of which are scored during an
actual examination, were not included as part of this research. In addition, the impact and
interactions among and between the gender and race facets could not be estimated
because of confounding of the variables due to characteristics of the candidates depicted
in the simulation.

Implications

Given the critical nature of the decisions that are based on CPNE scores, it is
important to continually add to the body of validity evidence for the use of scores. In
addition, it is important to address misconceptions about clinical performance assessment
by educating stakeholders about the psychometric properties of these types of
examinations. With more data available about the psychometric properties of the CPNE,
nursing educators can refine the precision and dependability of the measurement. In
addition, psychometric data can serve as ammunition, in the form of evidence, for educators in support of competency-assessment models of education for nursing and other healthcare professions. Making validity evidence and threats to validity for high-stakes examinations fully transparent to stakeholders is an important aspect of ensuring the sustainability and ongoing credibility of non-traditional, competency-based models of education. In addition, such evidence may be useful in enhancing the legal defensibility of CPNE outcomes.

**Conclusions and Future Research**

This study was designed to identify the dependability and accuracy of CPNE scores. The findings suggest that CPNE scores are dependable estimates. At the same time, because of the high-stakes nature of the examination, it points to the need for increased and perhaps more frequent rater training to enhance scoring accuracy for some aspects of the scored items. Despite its limitations and delimitations, the study provides sound data to support the interpretive argument for the use of CPNE scores.

Future research should address the potential fatigue effect of raters by counterbalancing the participants into groups, scoring candidates in various orders (i.e., by having two rater groups with one group scoring candidate one first and one group scoring candidate two first). Also, adding the critical elements that involve scoring the recording of patient data would create a richer picture of the accuracy and dependability of scores.

The computer-delivered simulation aspect of the design may have impeded participation rates. The response rate for this research was well below Nettleton’s (1996) study (i.e., 25% versus 75%) that employed a face-to-face approach, with raters scoring
the video simulation on televisions in a group setting at various sites across the country with the researcher present. This is an important consideration in designing future research in a manner that optimizes participation rates. While the response rate yielded adequate power for the types of data analysis conducted, increased participation rates would have been desirable and may have had a couple of benefits. One potential benefit is that it may have yielded a more representative sample in terms of months in the rater role. In addition, while no population data was available for some of the demographic characteristics such as ethnicity and full-time work role, increased participation rates would create a better potential for yielding a more comparable sample in these areas.

In addition, future research should be aimed at teasing out other facets to include in additional G-study analysis, including specific rater, candidate, and item characteristics as well as perhaps characteristics of the environments where the examination is administered. Using data from the G-study along with other available validity evidence, a D-study could be employed to further improve the design of the CPNE. A D-study would allow the assessor to determine how the CPNE would change if an aspect or facet were altered, thereby determining the optimal conditions under which the CPNE would produce the most dependable scores (Benner, 2001).

Finally, candidate selection in terms of ethnic background is an important consideration for future research designs so that comparisons can be made. Considering candidate ethnicity in future designs is particularly important because prior G-study analysis focusing on the stability and dependability of CPNE scores revealed variability in CPNE outcomes between ethnic groups (Yarbrough, et al., 2007).
References

American Educational Research Association, American Psychological Association, &
and psychological testing*. Washington, DC: American Educational Research
Association.

Prentice Hall.


Downing, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 711-


Hopkins University Press.

Bosco, C. (2003). *CPNE Inter-rater reliability study*. Written report to Nursing Faculty
Committee. Excelsior College, Albany, NY.

(Ed.), *New directions for testing and measurement: Methodological developments*
(*No. 4*). San Francisco: Jossey-Bass.


The Regents of the University of California. retrieved from ERIC, December 30, 2008.


Clinical Performance in Nursing Examination Technical Manual: Sources of
validity evidence and threats, Excelsior College.
Appendix A: Pilot #1: Measurement of the Fidelity of the Simulation Video

Six Excelsior College current and former associate degree program faculty members and consultants who were deemed unbiased but informed experts viewed the first version of the simulation video and responded to the following 11 items measuring various aspects of the fidelity of the simulation in representing an actual CPNE. Nine of the item responses were measured using the following 7-point Likert scale (i.e., items 1, 2, 3, 4, 5, 6, 7, 9, 10):

1. strongly disagree
2. disagree
3. somewhat disagree
4. neutral
5. somewhat agree
6. agree
7. strongly agree

Items 2 and 4 were worded negatively and scored accordingly. Two items (i.e., 8 & 10) required open ended responses requesting recommendations for changes to the scoring tool and orientation script.

The items:

1. The video accurately depicts the type of verbal interactions between a candidate and a patient during an actual CPNE administration.
2. The video does not accurately represent the behaviors that a candidate exhibits for the areas of care encountered during an actual CPNE administration.
3. The video accurately portrays the behaviors that a patient exhibits during an actual CPNE administration.
4. The video does not accurately depict the behaviors that a clinical examiner demonstrates during an actual CPNE administration.
5. The video accurately represents the type of environment where the CPNE is administered.
6. The CPNE Research Scoring Tool conforms to the scoring form used by clinical examiners during an actual CPNE.
7. The CPNE Research Scoring Tool is user-friendly.
8. Please share any recommendations you have for edits to the *CPNE Research Scoring Tool*.

9. The orientation to the research at the beginning of the video provides clear and sufficient information for study participants.

10. Please share any recommendations you have for additional or clarifying orientation information necessary for study participants.

11. Areas of care assigned to each student depicted in the video are of comparable difficulty.

Results

Table 1 contains means, range of scores, and standard deviations for each item measuring fidelity of the simulation video.

**Table 1. Descriptive Statistics for Fidelity Items**

<table>
<thead>
<tr>
<th>Item</th>
<th>N</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Verbal Interactions</td>
<td>6</td>
<td>6.00</td>
<td>7.00</td>
<td>6.3333</td>
<td>.51640</td>
</tr>
<tr>
<td>2. Candidate Behaviors</td>
<td>6</td>
<td>1.00</td>
<td>2.00</td>
<td>1.6667</td>
<td>.51640</td>
</tr>
<tr>
<td>3. Patient Behaviors</td>
<td>6</td>
<td>6.00</td>
<td>7.00</td>
<td>6.3333</td>
<td>.51640</td>
</tr>
<tr>
<td>4. Examiner Behaviors</td>
<td>6</td>
<td>1.00</td>
<td>3.00</td>
<td>1.8333</td>
<td>.75277</td>
</tr>
<tr>
<td>5. Environment</td>
<td>6</td>
<td>6.00</td>
<td>7.00</td>
<td>6.1667</td>
<td>.40825</td>
</tr>
<tr>
<td>6. Scoring Tool Conforms</td>
<td>6</td>
<td>3.00</td>
<td>7.00</td>
<td>5.6667</td>
<td>1.36626</td>
</tr>
<tr>
<td>7. Scoring Tool User-friendly</td>
<td>6</td>
<td>5.00</td>
<td>7.00</td>
<td>6.0000</td>
<td>.63246</td>
</tr>
<tr>
<td>8. Orientation Clear and Sufficient</td>
<td>6</td>
<td>6.00</td>
<td>7.00</td>
<td>6.1667</td>
<td>.40825</td>
</tr>
<tr>
<td>9. Comparable Difficulty</td>
<td>6</td>
<td>3.00</td>
<td>7.00</td>
<td>5.1667</td>
<td>1.72240</td>
</tr>
<tr>
<td>Valid N (listwise)</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

All of the items scored using the Likert scale were acceptable to the researcher. The *somewhat agree* response to item 6 was expected because there were modifications to the CPNE Scoring Tool for the research project due to the fact that only particular aspects of the examination were being scored and having all the actual examination critical elements on the form would have had a high potential to confuse the research participants. Thus, only scored areas of care and critical elements were included on the form. The *somewhat agree* response was primarily due to the fact that some of the experts viewed the first candidate’s assigned areas of care more slightly more difficult than the second candidate’s assigned areas of care because both of the areas assigned to candidate one are considered “management” areas of care as opposed to the “assessment” areas of care assigned to candidate two. The variance in the candidate assignments was deliberated during the script development phase by three faculty experts who deemed the assignments comparable due to the nature of the areas of care and the fact that the first
candidate was assigned two selected areas of care and the second candidate was assigned three selected areas of care.

All experts found the orientation to the research depicted on the video thorough and clear. No recommendations were made for edits to the orientation script. Experts did recommend changes to the scoring tool, including more demarcation between areas on the form delineated for each candidate, eliminating all recording critical elements from the form, and changing the order of areas of care depicted on the form for candidate number 2. The experts also recommended two changes to the video including insertion of additional video clearly capturing the first student washing her hands before going to the medication cart and very deliberate observation of the abdomen for candidate number 2.
Appendix B: Demographic Data and Scoring Tool

Demographic Data

RPAC Affiliation (choose primary RPAC affiliation)

MPAC    NPAC    SPAC

Primary Test Site Hospital Affiliation(s) ____________________________________________

Age

Gender (M = male; F = female)

Please answer both questions (a) and (b) below.

(a) Are you Hispanic or Latino/Latina? Yes No

(b) Please select the racial/ethnic group(s) you identify with regardless of your answer to the
above question (you may select more than one):

American Indian or Alaska Native    Black or African American    White
Asian                               Native Hawaiian or Other Pacific Islander

Highest degree earned (choose one)

MS/MA    PhD or EdD

Time in Clinical Examiner Role

Years    Months

Full-time Work Role (choose one)

Educator
Staff Development
Administrator—Education
Administrator—Acute or Long-Term Care Facility
Administrator—Other
Nurse Practitioner
Clinical Nurse Specialist
Other ________________________________________________________________

How many times have you worked as a Clinical Examiner or Clinical Associate within the past
12 months?
CPNE Research Scoring Tool

Student #1

Clinical Decision Making

The problem solving process by which choices are made in nursing practice. This process involves the identification of a patient problem, selection of a course of action or nursing intervention in response to a patient situation and an evaluation of a patient's progress that is based on theory, scientific principles, established protocols, and information presented in pertinent references. Clinical Decision Making as it is defined in the study guide, is demonstrated in the CPNE® during all phases of the nursing process. In the Implementation Phase clinical decisions are observed through the implementation of critical elements. However, when a student makes a deliberate decision to omit or modify a critical element, the reason for the omission or modification must be verbalized to the Clinical Examiner at the time of the omission or modification. The Clinical Examiner and/or Clinical Associate will determine the acceptability of that decision. An incorrect decision results in failure of the Patient Care Situation.

Area of Care: ________________________________________________________________

Critical Element(s) Omitted: __________________________________________________

Student’s statement of reason for omission or modification for one or more critical elements:

__________________________________________________________________________

__________________________________________________________________________

__________________________________________________________________________

__________________________________________________________________________

Overriding Areas of Care

Asepsis

The prevention of the introduction and/or transfer of microorganisms. Special consideration should be given to handwashing before and during each PCS as required by principles of asepsis. Any time a violation of asepsis occurs, the entire PCS will be terminated and failed.

The successful student:

1. Washes hands in the presence of the Clinical Examiner before beginning the Implementation Phase of each PCS

86
Caring

A pattern of behaviors that pervades the nurse-patient interaction as characterized by attentiveness to others’ experiences, the establishment of a trusting relationship with the patient and/or significant other, and respect for the values, dignity and culture of others.

The successful student:

1. Establishes communication and a trusting relationship with the patient at the beginning of the Implementation Phase by:
   a. Introducing self

   and

   b. Identifying the patient by comparing two of the following pieces of patient information against the Student PCS Response Form Kardex:
      1) Patient name
      2) Date of Birth
      3) Medical record number

   and

   c. Explaining the purpose of the interaction
<table>
<thead>
<tr>
<th>or</th>
</tr>
</thead>
<tbody>
<tr>
<td>d. Using touch with a patient who is a child or noncommunicating</td>
</tr>
<tr>
<td>adult if culturally appropriate</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>2. Uses therapeutic communication techniques consistent with the</td>
</tr>
<tr>
<td>patient’s level of understanding to interact with the patient and</td>
</tr>
<tr>
<td>significant others by:</td>
</tr>
<tr>
<td>a. Encouraging the patient’s expression of needs</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>b. Responding to the patient’s verbal expressions</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>c. Responding to the patient’s nonverbal expressions</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>d. Facilitating goal-directed interactions by:</td>
</tr>
<tr>
<td>1) Explaining the nursing actions to be taken</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>2) Asking questions to determine the patient’s response to nursing</td>
</tr>
<tr>
<td>care</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>3) Asking questions to determine the patient’s comfort level</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>4) Focusing communication toward patient-oriented interests</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>5) Eliciting the patient’s choices/desires in the organization of</td>
</tr>
<tr>
<td>care</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>3. Uses verbal expressions that are not overly familiar, patroniz</td>
</tr>
<tr>
<td>ing, demeaning, abusive, or otherwise unacceptable</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>4. Uses physical expressions that are not overly familiar, patroni</td>
</tr>
<tr>
<td>zing, demeaning, abusive, or otherwise unacceptable</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>5. Relates in a manner that respects the values, dignity, and cul</td>
</tr>
<tr>
<td>ture of others</td>
</tr>
</tbody>
</table>

Critical Element(s) Failed: ____________________________________________

Description of student behavior relative to specific element failed:

____________________________________________________________________
____________________________________________________________________
____________________________________________________________________
____________________________________________________________________
____________________________________________________________________

88
__ Emotional Jeopardy

Any action or inaction on the part of the student which threatens the emotional well-being of the patient or significant others. This area is invoked at the discretion of the Clinical Examiner, validated with the patient, and supported by data from the clinical situation. The entire PCS will be terminated and failed any time the emotional well-being of the patient or significant other is threatened.

Description of student behavior relative to failure for emotional jeopardy:

__________________________________________________________________________
__________________________________________________________________________
__________________________________________________________________________
__________________________________________________________________________

__ Physical Jeopardy

Any action or inaction on the part of the student which threatens the patient’s physical well-being. Students are accountable for the patient’s safety throughout the entire PCS. Any time the physical safety of the assigned patient is threatened through omission, such as not reporting a deterioration in the patient’s clinical condition, or by imminent incorrect action by the student, the entire PCS will be terminated and failed. This area of care is to be invoked at the discretion of the Clinical Examiner and supported by data from the clinical situation.

Description of student behavior relative to failure for physical jeopardy:

__________________________________________________________________________
__________________________________________________________________________
__________________________________________________________________________
__________________________________________________________________________

Note: Mobility will not be scored.

Selected Areas of Care

__ Respiratory Management

The assessment of respiratory status and the encouragement of, instruction about, assistance with, and the determination of the effectiveness of respiratory hygiene activities. Respiratory hygiene activities include deep breathing, coughing, chest percussion, suctioning, and/or the use of mechanical devices. (Respiratory Management will not be assigned in the same PCS with Respiratory Assessment.)
The successful student:

1. Complies with established guidelines

2. Positions the patient to facilitate respiratory hygiene activity(ies)

3. Provides a receptacle to receive secretions as needed

4. Assesses the patient’s respiratory status *before* initiating respiratory hygiene activity(ies) by:
   a. Instructing the patient specifically to breathe in and out as deeply as possible
   b. Auscultating breath sounds over upper and lower lobes by systematically moving the stethoscope from side to side
   c. Observing breathing patterns

5. Directs the patient or performs in one or more respiratory hygiene activity(ies)
   a. Deep breathing:
      1) Instructs the patient specifically to breathe in and out as deeply as possible
      2) Repeats deep breathing exercise as ordered or as indicated by the patient’s condition
   b. Coughing:
      1) Instructs the patient specifically to breathe in and out deeply
      2) Instructs the patient specifically to cough forcefully *on third or fourth expiration*
      3) Provides for splinting, while the patient is coughing, if necessary
   c. Mechanical devices, such as those used for inspiratory spirometry, etc.:
      1) Instructs the patient specifically to use the device
      2) Repeats respiratory exercise as ordered or as indicated by the patient’s condition
   d. Chest Percussion:
      1) Claps the designated area(s) of the chest wall vigorously with cupped hands, *unless contraindicated*
      2) Vibrates the designated area(s) of the chest wall vigorously *unless contraindicated*
   e. Suctioning:
      1) When suctioning by catheter is assigned:
<table>
<thead>
<tr>
<th>a)</th>
<th>Verifies patency of the catheter</th>
</tr>
</thead>
<tbody>
<tr>
<td>b)</td>
<td>Sets the pressure on the suction machine as designated</td>
</tr>
<tr>
<td>c)</td>
<td>Inserts the catheter <em>before</em> suctioning</td>
</tr>
<tr>
<td>d)</td>
<td>Rotates the catheter <em>continuously during</em> suctioning</td>
</tr>
<tr>
<td>e)</td>
<td>Suctions for <em>no more then 15 seconds</em> at a time</td>
</tr>
<tr>
<td>f)</td>
<td>Repeats as necessary to remove secretions</td>
</tr>
</tbody>
</table>

or

<table>
<thead>
<tr>
<th>2)</th>
<th>When suctioning by bulb syringe is assigned:</th>
</tr>
</thead>
<tbody>
<tr>
<td>a)</td>
<td>Deflates the bulb syringe prior to insertion</td>
</tr>
<tr>
<td>b)</td>
<td>Inserts the bulb syringe into the patient’s mouth and/or nares <em>before</em> suctioning</td>
</tr>
<tr>
<td>c)</td>
<td>Aspirates secretions</td>
</tr>
<tr>
<td>d)</td>
<td>Repeats as necessary to remove secretions</td>
</tr>
</tbody>
</table>

6. Reassesses respiratory status *immediately after* respiratory hygiene activities

Critical Element(s) Failed: ____________________________________________________________

Description of student behavior relative to specific element failed:

_____________________________________________________________________________
_____________________________________________________________________________
_____________________________________________________________________________
_____________________________________________________________________________
_____________________________________________________________________________
Medications

The administration of medications by any route: oral, intramuscular, intravenous, subcutaneous, or other routes. (Must be completed successfully at least once during the CPNE®.)

The successful student:

1. Complies with established guidelines related to medication administration

2. Selects the prescribed medication using the hospital medication administration record (MAR)

3. Measures the prescribed dosage

4. Identifies the patient *immediately before* administering the medication by comparing two of the following pieces of patient information against the MAR:
   a. Patient name
   b. Date of birth
   c. Medical record number

5. Uses the correct needle size for injections

6. Uses the prescribed route and/or site for administering medications

7. Administers the prescribed medications to the designated patient

8. Administers the medication within 30 minutes of the scheduled time

9. When IV medication is to be administered:
   a. Records the correct flow rate in drops per minute for gravity flow or milliliters per hour for infusion control devices (ICDs) on the PCS Recording Form before administering the medication
   b. Assesses the insertion site for dislocation, infiltration, or other complications *immediately before* administering the medication by using one of the following methods:
      1) Feeling the surrounding skin for changes in temperature
      2) Palpating the surrounding tissue for edema
c. Clears air from tubing before initiating flow

d. When an intermittent venous access device is used:
   1) Aspirate for blood return unless contraindicated
   2) Flushes with the designated solution *prior* to medication administration
   3) Flushes with the designated solution *after* medication administration
   4) Records the flush solution used on PCS Response Form

e. Regulates the flow to deliver the prescribed amount in the designated period of time
   (± 5 drops per minute for gravity flow or the correct ICD setting)

Critical Element(s) Failed: __________________________________________________________

Description of student behavior relative to specific element failed:

________________________________________________________________________________
________________________________________________________________________________
________________________________________________________________________________
________________________________________________________________________________
________________________________________________________________________________

Comments Related to Student Performance and Scoring

________________________________________________________________________________
________________________________________________________________________________
________________________________________________________________________________
________________________________________________________________________________

Student #2

---

Clinical Decision Making

The problem solving process by which choices are made in nursing practice. This process involves the identification of a patient problem, selection of a course of action or
nursing intervention in response to a patient situation and an evaluation of a patient’s progress that is based on theory, scientific principles, established protocols, and information presented in pertinent references. Clinical Decision Making as it is defined in the study guide, is demonstrated in the CPNE® during all phases of the nursing process. In the Implementation Phase clinical decisions are observed through the implementation of critical elements. However, when a student makes a deliberate decision to omit or modify a critical element, the reason for the omission or modification must be verbalized to the Clinical Examiner at the time of the omission or modification. The Clinical Examiner and/or Clinical Associate will determine the acceptability of that decision. An incorrect decision results in failure of the Patient Care Situation.

Area of Care: ____________________________________________________________

Critical Element(s) Omitted: ______________________________________________

Student’s statement of reason for omission or modification for one or more critical elements:
_____________________________________________________________________
_____________________________________________________________________
_____________________________________________________________________
_____________________________________________________________________
_____________________________________________________________________

Overriding Areas of Care

_ Asepsis

The prevention of the introduction and/or transfer of microorganisms. Special consideration should be given to handwashing before and during each PCS as required by principles of asepsis. Any time a violation of asepsis occurs, the entire PCS will be terminated and failed.

The successful student:

1. Washes hands in the presence of the Clinical Examiner before beginning the Implementation Phase of each PCS
2. Protects self, others, and the environment from contamination
3. Protects the patient from contamination
4. Disposes of contaminated material in the designated container(s)
5. Establishes a sterile field when required

Critical Element(s) Failed: ______________________________________________________________

Description of student behavior relative to specific element failed:

________________________________________________________________________________

________________________________________________________________________________

________________________________________________________________________________

________________________________________________________________________________

________________________________________________________________________________

Caring

A pattern of behaviors that pervades the nurse-patient interaction as characterized by attentiveness to others’ experiences, the establishment of a trusting relationship with the patient and/or significant other, and respect for the values, dignity and culture of others.

The successful student:

<table>
<thead>
<tr>
<th>1. Establishes communication and a trusting relationship with the patient at the beginning of the Implementation Phase by:</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Introducing self</td>
</tr>
<tr>
<td>b. Identifying the patient by comparing two of the following pieces of patient information against the Student PCS Response Form Kardex:</td>
</tr>
<tr>
<td>1) Patient name</td>
</tr>
<tr>
<td>2) Date of Birth</td>
</tr>
<tr>
<td>3) Medical record number</td>
</tr>
<tr>
<td>and</td>
</tr>
<tr>
<td>c. Explaining the purpose of the interaction</td>
</tr>
<tr>
<td>or</td>
</tr>
<tr>
<td>d. Using touch with a patient who is a child or noncommunicating adult if culturally appropriate</td>
</tr>
</tbody>
</table>

2. Uses therapeutic communication techniques consistent with the patient’s level of understanding to interact with the patient and significant others by:

| a. Encouraging the patient’s expression of needs |
### Emotional Jeopardy

Any action or inaction on the part of the student which threatens the emotional well-being of the patient or significant others. This area is invoked at the discretion of the Clinical Examiner, validated with the patient, and supported by data from the clinical situation. The entire PCS will be terminated and failed any time the emotional well-being of the patient or significant other is threatened.
Description of student behavior relative to failure for emotional jeopardy:

_________________________________________________________________________________
_________________________________________________________________________________
_________________________________________________________________________________
_________________________________________________________________________________
_________________________________________________________________________________

__ Physical Jeopardy

Any action or inaction on the part of the student which threatens the patient's physical well-being. Students are accountable for the patient's safety throughout the entire PCS. Any time the physical safety of the assigned patient is threatened through omission, such as not reporting a deterioration in the patient's clinical condition, or by imminent incorrect action by the student, the entire PCS will be terminated and failed. This area of care is to be invoked at the discretion of the Clinical Examiner and supported by data from the clinical situation.

Description of student behavior relative to failure for physical jeopardy:

_________________________________________________________________________________
_________________________________________________________________________________
_________________________________________________________________________________
_________________________________________________________________________________

Note: Mobility will not be scored.

Selected Areas of Care

__ Peripheral Vascular Assessment

The assessment of temperature, perfusion, pulse, sensation, and movement in patients with casts, traction, or peripheral vascular impairment. When possible, this assessment would include a comparison of extremities.

The successful student:

1. Complies with established guidelines

2. Compares the extremities by all of the following:
   a. Palpating for the presence or absence of the most distal pulses
b. Comparing the most distal corresponding palpable pulses

c. Assessing perfusion of extremity(ies) by:
   1) Checking capillary refill

or
   2) Observing color

d. Assessing for temperature of extremity(ies)

e. Eliciting the patient’s response to tactile stimuli applied to the distal portion of the extremity(ies)

f. Assessing motor function by:
   1) Asking the patient to move extremity(ies)

or
   2) Noting movement of the extremity(ies) in a child under 3 or a noncommunicating adult

Critical Element(s) Failed: __________________________________________________________

Description of student behavior relative to specific element failed:
_____________________________________________________________________________
_____________________________________________________________________________
_____________________________________________________________________________
_____________________________________________________________________________

Neurological Assessment

The assessment of neurological status including level of consciousness, equality of pupil size and reaction to light, sensorimotor responses, and palpation of the anterior fontanel in a child under 1 year of age.

The successful student:

1. Complies with established guidelines

2. Assess the patient’s level of consciousness by:
   a. Asking specific questions to determine orientation to all of the following:
      1) Time
      2) Place
| 3) Person
or  
    b. Determining the patient’s ability to recognize familiar people or common objects in the environment
or
  
    c. Presenting visual, auditory, and tactile stimuli to a child between 1 and 3 years of age or a noncommunicating child or adult

3. Palpates the anterior fontanel of a child under 1 year of age, with the child in an upright position, unless contraindicated

4. Assesses pupillary response regarding:
   a. Equality of pupil size

   and

   b. Reaction to light

5. Assesses equality of the motor response in upper and lower extremities in a responsive patient by:
   a. Asking the patient to:
      1) Use both hands to squeeze the student's hands simultaneously

      and

      2) Dorsiflex or plantarflex both feet simultaneously against resistance

   or

   b. Observing musculoskeletal response(s) in a child under 3 years of age or a noncommunicating child or adult for:
      1) Symmetry

      and

      2) Movement

6. Assesses the patient's response to a noxious stimulus when the patient is nonresponsive to verbal stimuli by applying pressure to a nailbed

Critical Element(s) Failed: __________________________________________________________

Description of student behavior relative to specific element failed:

____________________________________________________________________________
____________________________________________________________________________
____________________________________________________________________________
____________________________________________________________________________
____________________________________________________________________________
____________________________________________________________________________
Abdominal Assessment

The inspection, auscultation, light palpation, and measurement of the abdomen for the presence of bowel sounds, distention, rigidity, and tenderness.

The successful student:

1. Complies with established guidelines
2. Positions the patient to facilitate abdominal assessment
3. Inspects for distention
4. Auscultates for bowel sounds over all 4 quadrants
5. Performs light palpation over all 4 quadrants for tenderness or rigidity, unless contraindicated

Critical Element(s) Failed: 

Description of student behavior relative to specific element failed:

Comments Related to Student Performance and Scoring
Appendix C: Data Analysis with Outlier Included: Candidate Two

Percent Accuracy of Rater Scoring

![Pie Chart](chart.png)

ANCOVA for Rater Accuracy in Scoring

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
<th>Partial Eta Squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corrected Model</td>
<td>12.353</td>
<td>45</td>
<td>.275</td>
<td>1.568</td>
<td>.133</td>
<td>.771</td>
</tr>
<tr>
<td>Intercept</td>
<td>.051</td>
<td>1</td>
<td>.051</td>
<td>.291</td>
<td>.595</td>
<td>.014</td>
</tr>
<tr>
<td>Number of Times worked in the past year</td>
<td>.123</td>
<td>1</td>
<td>.123</td>
<td>.704</td>
<td>.411</td>
<td>.032</td>
</tr>
<tr>
<td>Number of months in the role of rater</td>
<td>12.048</td>
<td>44</td>
<td>.274</td>
<td>1.564</td>
<td>.135</td>
<td>.766</td>
</tr>
<tr>
<td>Error</td>
<td>3.677</td>
<td>21</td>
<td>.175</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>22.000</td>
<td>67</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>16.030</td>
<td>66</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ R^2 = .771 \text{ (Adjusted } R^2 = .279) \]

Analysis of covariance in rater scoring of candidate two with times worked and months in role analyzed as covariates revealed that neither covariate was significantly
related to the rater accuracy in scoring $F(1, 21) = .704, \ p<.05$ and $F(44, 21) = 1.564, \ p<.05$. 
June 14, 2010

Dear Laurie Nagelsmith,

Taylor & Francis hereby grants you permission to re-print the figure which appears on page 628 from *Handbook of Test Development* 9780805852653 in a dissertation for SUNY Albany in 2010.

Permission is given on a one-time, *nonexclusive* basis. Future uses of the material must be applied permission for.

If you need further permission than this, let me know!

Best Regards,

Mindy Rosenkrantz
Permissions Coordinator
Taylor and Francis Group
270 Madison Avenue 4th Floor
New York, NY  10016
Permissions.Mailbox@taylorandfrancis.com