©Copyright 2007
Kirsten Colleen Roberts

A Validity and Reliability Study of the Objective Structured Clinical Examination

Kirsten Colleen Roberts

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2007

Program Authorized to Grant Degree: College of Education

UMI Number: 3275905

Copyright 2007 by
Roberts, Kirsten Colleen

All rights reserved.

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

UMI®

UMI Microform 3275905
Copyright 2007 by ProQuest Information and Learning Company.
All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

University of Washington
Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

Kirsten Colleen Roberts

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Chair of the Supervisory Committee:

Catherine Taylor

Reading Committee:

Catherine Taylor

Robert Abbott

Douglas Brock

Date: June 7, 07

In presenting this dissertation in partial fulfillment of the requirements for the doctoral degree of at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of the dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to ProQuest Information and Learning, 300 North Zeeb Road, Ann Arbor, MI 48106-1346, 1-800-521-0600, to whom the author has granted "the right to produce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform."

Signature Kirsten Roberts

Date 6-7-07

University of Washington

Abstract

A Validity and Reliability Study of the Objective Structured Clinical Examination

Kirsten Colleen Roberts

Chair of the Supervisory Committee
Associate Professor Catherine S. Taylor
Educational Psychology

Medical education is seeing a renewed emphasis on clinical skills proficiency. Unlike medical knowledge assessment, clinical skills cannot be adequately assessed using paper-and-pencil tests. The Objective Structured Clinical Examination (OSCE) is now generally accepted as a means of assessing medical students' clinical skills proficiency. Given the growing popularity of the OSCE, research is needed to support the validity and reliability of scores from these performance based assessments.

Data for this study were obtained from the 2005 and 2006 senior OSCE administrations at the University of Washington, School of Medicine. Scores were obtained for 346 students. The four medical cases common to both years were identified and used in this study. OSCE validity and reliability issues were explored using classical test theory principles and structural equation modeling techniques.

Several research questions were of interest. Are the OSCE checklist scores measuring a unitary dimension or multiple dimensions of clinical skills? The results of the structural analysis and a confirmatory factor analysis suggested that the OSCE

checklists were measuring multiple dimensions. A single overall rating was assigned to each medical case; however, a score for each dimension might have provided a more accurate assessment of the student's clinical skills and would have allowed for more effective formative instruction.

Are all items on the OSCE checklist equally predictive of final judgment about student proficiency? The data suggested that, in some cases, items from the write-up component of the medical case checklist appeared to have more influence on the overall score than did the items from the standardized patient encounter checklist.

How well do OSCE scores from a refined checklist predict other judgments of student proficiency? The scores from the revised subscales were correlated with USMLE Step 2 Clinical Knowledge (CK) scores for discriminant evidence of construct validity. Although several of the subscales did correlate with USMLE Step 2 CK scores at a statistically significant level, these correlations were not high enough to be valuable in a practical sense.

TABLE OF CONTENTS

List of Tables

Table Number Page

List of Figures

Figure Number Page

Acknowledgements

The author wishes to express sincere appreciation to her friends and family whose belief never wavered and who never failed to provide words of encouragement when they were most needed; to her colleagues at the UW School of Medicine for their tremendous support and encouragement; and to her advisor for being such a wonderful guide on this incredible adventure.

Dedication

To my mom and dad

Chapter 1: Introduction

Educational assessment methods and philosophies are constantly evolving. America has seen a slow and steady shift from individualized instruction and assessment in the early days of education, to standardized, pencil-and-paper testing, to the current trend of performance based assessment for many activities. A growing body of research in this area has fueled the movement. As with general testing and assessment, medical education testing and assessment is also evolving. Research studies are steadily emerging and contributing to this constant evolution. The following discussion will briefly describe the evolution of assessment, reasons for assessment, the need for and emergence of alternative forms of assessment, and the role of performance assessment in medical education.

Assessment in Education

Assessment has existed in one form or another for centuries. Formal oral exams have existed in universities since perhaps as early as 1219 at the University of Bolgna (DuBois, 1970). Paper testing was developed following the introduction of paper in the twelfth century and by the late sixteenth century written exams were in elementary schools. In universities in England, oral exams were in introduced in the 1600s and written exams in the 1800s (1970).

At the beginning of the twentieth century, behavioral studies were being influenced by science. Academicians and psychologists were turning to scientists of the time, believing that tests being used in the physical sciences could also be used to study mental operations (Giordano, 2005). This was applied in America during World War I,

when there arose a need for the psychological testing and scoring of many people in a short amount of time for the classification of military personnel (DuBois, 1970). It was here that the concept of the multiple-choice test item, a test question with two or more alternative answers, was first introduced.

In the early twentieth century there was growing dissatisfaction with the subjective nature of the oral format of examinations and the desire for more objective assessment increased (Giordano, 2005). The multiple-choice type question was one way to provide more objectivity. Although objective style assessments continued to gain in popularity, it was noted that the objectively scored, pencil-and-paper exam was more appropriate for some subjects, for instance mathematics, than for others (Giordano, 2005). Pencil-and-paper testing continued through the latter half of the 20th century, when great emphasis was placed on statewide testing as a means of determining academic achievement status (Jones, Jones, & Hargrove, 2003). As standardized, paper-and-pencil testing flourished, there also emerged a movement to explore alternative types of testing, such as performance based assessment, that would measure qualities other than those which can be assessed by a pencil-and-paper test. Now, when reference is made to alternatives to traditional testing, reference is really being made to alternatives to the standardized paper-and-pencil, multiple-choice types of tests (Messick, 1994).

Reasons for Assessment

Thoughts on the early purposes of educational assessment are many and varied. Some say that the original purpose of student assessment was to aid teachers in improving instruction in the classroom (McLean & Lockwood, 1996). Others suggest that

standardized tests were introduced as a means of assigning grades (Giordano, 2005). Still others say that achievement tests were originally intended for individual student evaluation (Jones, Jones, & Hargrove, 2003). Regarding higher education, Messick says, "For years the main function of assessment in the academy was selection to maximize the level of talent as an outcome of higher education" (1999, p. 3). Thoughts on the current purposes of assessment continue to vary. Berk (1986) says that the purpose of assessment is to differentiate among levels of performance. Messick (1999) lists instructional guidance and placement, career guidance and decision making, improvement of instruction and student performance, certification of learning and competence, and evaluation of program quality as reasons for educational assessment.

We also assess because the stakeholders demand it. Suskie states that, "The increased need for higher education has increased public attention to it. Higher education is no longer an optional indulgence but a necessity for economic well-being" (2006, p. 15). She also suggests that as long as education is viewed as an expensive endeavor, there will be significant interest from stakeholders and a corresponding interest in ensuring efficiency and effectiveness of the institution. Institutions responsible for education are accountable to myriad stakeholders: prospective and enrolled students and their parents, the institution's internal community, accrediting organizations, governments, the business community, mass media, taxpayers, and others (Dugan, 2006). The stakeholders want evidence that learning is occurring, and believe that "graduates should be able to communicate effectively, think critically, and solve problems, thereby raising the quality of the workforce to contribute to the economy and welfare of the locality and to compete

regionally, nationally, and globally" (Dugan, 2006, p. 40). In the broadest sense, it is this evidence of learning that educational assessment attempts to provide.

Assessment also promotes public safety through licensing and certification. As described in the *Standards for Educational and Psychological Testing* (AERA APA and NCME, 1999), state and local governments have imposed licensure requirements on many professions, including law, real estate, and medicine. Individuals wanting to obtain licensure in such an area must demonstrate that they have mastered the skills and knowledge associated with the particular domain and that minimum standards have been met. This is achieved through tests which often include combinations of multiple-choice type items, written essays, oral exams, and/or performance tasks. Accurate assessment in licensure assures the public of professionals who have mastered the requirements for the profession, and upholds the standards of the profession by assuring a minimum level of competency.

Given the broad purposes of assessment and the range of domains that are assessed, questions of assessment formats and the relationship between formats and validity have always been an area of research. However, since the invention of the multiple-choice test item, much of the research has been on the development of effective multiple-choice items to measure a broad range of knowledge and skills. More recent research has focused on alternatives to multiple-choice testing as a way to increase the validity of test scores.

Alternative Forms of Assessment

Significant changes in instruction in higher education are occurring. The lecture format still dominates instruction, but there is a shift towards involving the student as an active participant (Svinicki, 2005). Participation will generally require the student to complete tasks or demonstrate skills that elicit more complex cognitive processes than simply reciting a memorized statement of fact (AERA APA and NCME, 1999). Messick supports this as he believes that a "...likely hallmark of educational performance assessments is their nearly universal focus on higher-order thinking and problem-solving skills" (1994, p. 5).

This recognition of the need to include assessment of thought processing activities, and not only fact-driven knowledge, is an important development. Haladyna (1994) points out that it is now generally accepted that the focus on the measurement of higher order thinking will continue to be preeminent. He also acknowledges, however, the difficulties associated with assessing higher level thinking, suggesting that one can never be sure what mental process is truly being measured. As such, providing evidence of the validity of score inferences from assessments that are attempting to capture higher order thinking continues to be a challenge.

There are different types of performance assessment, for instance performance tasks, written scenarios, and portfolios (National Research Council, 2002), exhibitions and demonstrations, and classroom presentations (McLean & Lockwood, 1996). When forms of learning other than traditional didactics occur, alternate forms of assessment must also be implemented (Svinicki, 2005). Indeed, there are many areas in which

assessment can only be achieved through assessing performance, such as dance, athletics, or, as this research study will further explore, the clinical skills of physicians-in-training.

The notion of authentic assessment becomes important here. Authentic assessment is described as "based on student activities that replicate real-world performances as closely as possible" (Svinicki, 2005, p. 23). Wiggins (as cited in Svinicki, 2005) describes characteristics of authentic assessment as asking the student to "do," or perform, requiring judgment with more than one correct answer, and allowing for feedback, practice, and second chances. Messick (1994) proposes that authenticity implies the validity standard of construct representation or minimal construct under-representation. While there is great benefit in creating this approximation of a real world experience, there are also challenges associated with this type of assessment. Additional time, effort, and financial resources are invested, certainly; but there are also challenges from a measurement perspective (Svinicki, 2005). Obtaining evidence of score validity and reliability from performance assessments becomes key.

Assessment in Medical Education

Medical education in the United States, as we know it today, began with the Flexner report (1910). Recognizing the disorder of medical education, the inconsistent quality of physicians, and the importance of uniformity in creating quality physicians, the Flexner report outlined in detail how a medical school should look. The Flexner report set forth many guidelines, including the need for prerequisites for entrance into medical school, the requirement that didactics and lab work take place in the first two years of

medical school, and the requirement that the third and fourth years be spent in hospitals. The report sites many detailed standards, many of which are still in effect today.

With respect to medical student assessment, we find that developments in this particular area follow the trends in education in general. In the 1950s and 1960s, the traditional form of assessing medical knowledge was the essay. There were, however, significant concerns about the reliability of scoring as well as the limited range of knowledge that could be tested (Newble, 2004). Addressing these concerns, objectively scored tests (i.e. multiple-choice) rose in popularity and almost entirely replaced essays for assessing knowledge and recall (2004).

As with education in general, there are aspects of medical education that cannot be measured by pencil-and-paper testing alone and are better suited to alternative means, such as performance testing. Medical education not only requires a foundation of medical science knowledge, which can be assessed by paper-and-pencil tests, but also requires clinical skills competency, which calls for performance based assessment. It is well documented that traditional forms of assessment alone may not adequately capture clinical competence (Chirayu Auewarakul, Downing, Jaturatamrong, & Praditsuwan, 2005).

Harden, Stevenson, Downie, and Wilson (1975) acknowledge the wide use of objectively scorable techniques, such as multiple-choice questions, in knowledge testing. They further acknowledge that applying these techniques to clinical skills assessment is "impracticable" (p. 447). Until the 1970s, clinical skills were assessed by the clinical examination (1975). In this exam, the student encountered real patients in a hospital

setting. One or more physicians would observe the encounter and grade the student. This style of exam proved problematic in several ways. The patient encounters were highly variable and there was little consistency in marking standards between examiners. Additionally, there was confusion about what was being tested, as the physical exam depended on the patient, and may have resulted in simply a test of knowledge (1975). It was in response to these concerns that Harden introduced the objective clinical examination.

Harden described a process for assessing clinical skills designed to alleviate the problems associated with the clinical examination (1975). In his objective clinical exam, students rotated through a series of stations. They encountered a patient on whom the student was expected to perform a procedure. The encounter may have been scored by an evaluator using a checklist. The student was then required to answer a series of questions about the patient's case. This write-up was also scored by an evaluator. The student then received a score based on the number of points accumulated from the score sheets. With some modifications, Harden's objective clinical examination exists today much as he described it over thirty years ago. It is generally referred to as the Objective Structured Clinical Examination (OSCE) and is accepted globally for clinical skills assessment in medical education.

Harden was not alone in seeing the value of objective clinical examinations. The need to address the disconnect between the assessment of medical science knowledge using paper-and-pencil tests and the assessment of clinical skills needed for practice is reflected in the most recent modification to Step 2 of the United States Medical Licensing

Examination (USMLE). Beginning with the graduating class of 2005, medical students would be required to pass a Clinical Skills (CS) component in addition to the already existing Clinical Knowledge (CK) component. The National Board of Medical Examiners (NBME), the board that oversees the development of the USMLE, recognized that research suggested that poor communication and poor general clinical skills are related to medical malpractice suits, as well as to lower treatment compliance and lower patient satisfaction (USMLE, 2004). The NBME recognized the differences between the cognitive skills and knowledge required to understand diseases, and the clinical skills required to diagnose diseases, treat patients, and consult with colleagues. Further, it recognized that multiple-choice tests, such as the USMLE Step 2 CK, cannot by themselves adequately assess clinical skills. Students may pass a multiple-choice exam, yet lack the clinical skills necessary to practice medicine.

To address the inconsistencies among medical schools' teaching and assessment of clinical skills, and to establish a national standard, the board added the USMLE Step 2 CS component in 2004 (USMLE, 2004). This component of the exam is based on objective, structured, performance based methods of clinical skills assessment. The student rotates through 12 stations, where he or she encounters an actor playing the role of a patient. The student examines the patient and then writes up the patient history, physical exam findings, and a follow-up plan for further evaluation. The student receives a score of either "pass" or "fail" for each station (USMLE, 2007).

Other organizations have also recognized the importance of clinical skills. The Medical Council of Canada has incorporated clinical skills assessments into its Canadian

Licensing Examination. A clinical skills component was included in the certification process of the Education Commission for Foreign Medical Graduates; however, further recognizing the importance of clinical skills, foreign medical students are now expected to pass the USMLE Step 2 CS component before being allowed to enter residencies in the United States (Education Commission for Foreign Medical Graduates, 2007).

Through these changes in the USMLE, although knowledge based assessment retains its importance, higher level thinking and performance based assessment have emerged as equally important. Medical education is an active contributor to the growing body of literature on this subject as it attempts to respond to the renewed commitment to clinical skills competency in the physicians of the future. Creating performance based assessments that produce scores that are reliable and valid for their intended use can be challenging. It is expected that reliability and validity issues will continue to be prominent topics for future research in general education and, more specifically, in medical education, as researchers seek ways to accurately assess skills and abilities.

Purpose and Focus of this Research

Given the renewed emphasis on clinical skills proficiency, the desire to standardize the assessment of clinical skills, and the recent addition of the USMLE Step 2 CS component, research is needed to support the validity and reliability of scores from these medical education performance based assessments. Evidence is needed to support the assertion that the score interpretations from performance based assessments are accurate representations of the clinical skills proficiency of the examinees. The purpose of this research study was to further understand the reliability and validity issues

surrounding a specific clinical skills assessment instrument, the OSCE, by looking at the technical quality of the instrument at the item level. This research will explore the following questions: What is the internal consistency of checklist scores on different OSCE stations? Are all items on the OSCE checklist for a standardized patient encounter equally predictive of final judgment about student proficiency? How well do OSCE scores from a refined checklist predict judgments of student proficiency?

Significance of this Research

This study contributes to the existing body of literature regarding the effectiveness of the OSCE in assessing medical student clinical skills competency. It furthers the understanding of the reliability and validity issues surrounding a specific clinical skills assessment instrument. Furthermore, it applies the use of a measurement theory that a review of the literature would suggest is not often applied to explore the internal structure of this type of measurement instrument at the item level.

Organization of the Document

The remainder of this document is organized in the following manner. Chapter 2 reviews the existing literature regarding the three key points of the research study. First, the concept of validity that guided the research study is explored. Messick's concept of unified construct validity is the now generally accepted approach to validity, and the sources of evidence of construct validity applicable to this study are discussed. Second, the literature that already exists regarding the reliability and validity of the assessment instrument of interest for this study, the OSCE, is presented. Last, the measurement

approaches that guided the data analyses, classical test theory and confirmatory factor analysis, are reviewed.

Chapter 3 describes the particular research methods employed for the study. It describes how the scores were obtained and from whom. It describes the OSCE from which the scores were obtained, including a description of the medical cases and the types of items that comprised those medical cases. Last, it describes the analysis process and how classical test theory and structural equation modeling principles were applied to explore the OSCE data. Chapter 4 provides the results of the research. Finally, Chapter 5 provides a discussion of the results of the research, the limitations of the study, and possible directions for future research.

Chapter 2: Review of the Literature

This review of the literature focuses on the key elements of the study. First, the prevailing concept of validity, Messick's concept of unified construct validity, is described. The six sources of evidence of score validity are reviewed, followed by a discussion of threats to validity, and the evidence that is needed to help counteract those threats. Second, this review of the literature explores what is already known about the reliability and validity of Objective Structured Clinical Examination (OSCE) scores. Finally, the analyses used in the present study to explore the OSCE data are reviewed: classical test theory and confirmatory factor analysis, with particular attention to their application to score reliability and validity, and their application in test construction.

Validity

Although Messick's contributions are relatively recent, the idea that validity is preeminent among psychometric concepts has remained constant (Angoff, 1988). During the first fifty years following the introduction of the concept of validity, Angoff indicates that the concept received very little attention. He describes the general thrust in the 1930s as validity defined by the concept of correlation. This concept persisted into the 1940s with the extension that validity was seen as correlation with predictive value, or criterion validity. Other "types of validity" that came about at this time were concurrent and face validity. Although concurrent and predictive would later be combined into criterion-related validity, the 1954 Standards clearly identified four "types" of validity: content, predictive, concurrent, and construct validity.

The introduction of construct validity precipitated the changes in the generally accepted concept of validity. Beginning with work by Cronbach and Meehl suggesting that construct validity is a process requiring many lines of evidence (as cited in Angoff, 1988), and continuing through the 1960s and 1970s with the views of Messick and others, the idea of a unified concept of validity began to take hold. Under the unified concept of validity, all "types of validity" are subsumed under the main header of construct validity. Focus shifted away from the measurement instrument itself and onto the examinees' responses to the test items, score-based inferences, and score interpretations. Despite this shift, the literature contains a great many studies that still refer to types of validity.

Messick's Unified Concept of Construct Validity

Messick's (1989) seminal work on the unified concept of construct validity is the culmination of a hundred years of thinking about validity, and is becoming the basis upon which validity is generally viewed today. Messick (1989) defines validity as, "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment" (p. 13). An essential distinction between the old ideas of "types of validity" (face, construct, criterion, etc.) and the unitary concept of validity is the move to seeking "sources of validity evidence" as opposed to identifying types of validity. Indeed, the most current S*tandards for Educational and Psychological Testing* has adopted the unified concept and states that, "Validity is a unitary concept" (AERA APA and NCME, 1999, p. 11). This concept applies appropriately to both paper-and-pencil based assessments and performance based assessments.

Messick's (1989) unified concept of construct validity provides the foundation of the approach to examining score validity in common use today. The concept is grounded in the seeking of evidence of score interpretation validity from six possible sources. These six sources are content, substantive, structural (internal), external, generalizability, and consequential. The present study sought evidence of OSCE score validity from content, substantive, structural, and external sources.

Although Messick identifies several ways in which traditional pencil-and-paper tests differ from performance tests, he argues that his unified concept of construct validity applies equally well to both (1994). Fitzpatrick and Morrison claim that there is no distinction between performance assessments and other classes of tests (as cited in Messick, 1992); therefore, performance assessments should be evaluated by the same validity criteria as are other assessments (Messick, 1992). "In effect, the six interrelated aspects of unified validity provide a general framework for the validation of all assessments including performance assessments" (1992, p. 4).

Content as a source of construct validity. Content validity evidence resides "in the relation between the test and the domain of application" (p. 41); it looks at how well the student's responses to an assessment reflect that student's knowledge of the content area. The content aspect of validity involves identifying the boundaries of the domain to be assessed, then assuring that the tasks and questions chosen for the assessment reflect those boundaries and are relevant to and representative of the domain. If an area of the domain is not represented, there is potential exposure to threat of construct under-representation; if a quality is being measured other than what is intended, there is threat

of construct-irrelevant variance. In either case the score interpretation could not be considered valid for its intended use. The present study considered content as a source of construct validity by exploring the possible domains being assessed in the OSCE.

Substantive sources of construct validity. With substantive sources of construct validity evidence, the content aspect is expanded upon to include a requirement of empirical evidence. Items for the final assessment instrument are selected based on *consistent empirical responses to the tasks*. The evidence moves beyond professional judgment of content appropriateness by employing structural equation modeling techniques, such as factor analysis. In the case, of factor analysis, how the tasks load on a factor provides support for the domain representation and validity of score interpretation. The present study explored substantive sources of construct validity through a confirmatory factor analyses of refined checklist subscales.

Structural (internal) sources of construct validity. The structural sources of construct validity evidence address the *relationships among responses to the tasks, items, or parts of the test*. The internal structure of the test should be consistent with the internal structure of the domain. An item analysis, for example, may reveal that students with high overall scores all answered a particular item incorrectly. Likewise, students with low overall scores all answered that same item correctly. This should prompt an investigation of the item, as the relationship between the item response and the test responses as a whole is unexpected. The present study sought evidence of OSCE score validity by exploring the internal structure of the OSCE medical cases at the item level using classical test theory principles.

External sources of construct validity. The external sources of construct validity evidence explore the relationships between the assessment scores of interest and other measures and background variables of either the same or different constructs. Convergent evidence might be identified through high correlations between scores from tests of a similar construct, whereas discriminant evidence might be identified through low correlations with scores from tests of unrelated constructs. The present study sought evidence of OSCE score validity by exploring the relationships between the scores on OSCE medical cases and scores on another medical student assessment.

Generalizability as a source of construct validity. In one sense, generalizability refers to the differences in test processes and structures over time, across groups and settings, or in response to experimental interventions. A narrower set of tasks may provide greater reliability, but at the expense of generalizability across populations or settings. Likewise, a broader set of tasks may provide more generalizability, but at the expense of reliability. In another sense, generalizability applies to the transfer of tasks across the domain; in other words, how does performance on one task predict performance on another.

Consequences as a source of construct validity. This source of construct validity evidence is concerned with the social consequences of interpreting and using the test scores in particular ways. It is concerned with intended outcomes and unintended side effects. Any adverse impact on individuals or groups should not derive from any source of test invalidity.

There is no prescribed formula stating how many sources or how much evidence from each source is sufficient to substantiate a validity claim. Messick (1994) says that test score validity does not *depend on* any on particular source of validity, nor does it *require* any one particular source. Rather, one looks for a compelling argument that uses convergent and discriminant evidence to support the score interpretation. The *Standards for Educational and Psychological Testing* state that, "...a validity argument typically depends on more than one proposition..." and, "professional judgment guides decisions regarding the specific forms of evidence that can best support the intended interpretation and use" (AERA APA and NCME, 1999, p. 11).

Having described the sources of evidence of score validity, attention is now turned to potential threats to the validity of score inferences. These threats include construct under-representation and construct-irrelevant variance.

Potential Threats to Validity

Scores from performance based assessments are vulnerable to the same threats to validity as scores from traditional assessments. Messick (1989) identifies two such threats. The first, construct under-representation, occurs when the test lacks complete domain coverage and the assessment therefore fails to account for important dimensions of the construct. Take, for example, an OSCE that neglects to include items to test the student's ability to communicate well with a patient. The scores from such a test would not allow for a valid inference of the student's true level of clinical skill proficiency because a significant portion of the construct was omitted. Likewise, the USMLE Step 2

without the Clinical Skills component under-represented the domain of knowledge and skills needed to be licensed as a physician.

The second threat, construct-irrelevant variance, occurs when something is being assessed that distorts or interferes with the construct assessment. For example, an OSCE in which a patient-actor inadvertently provides "clues" to the student regarding the ailment or construct being tested. In this case, the student's score could not be interpreted as a valid reflection of his or her level of clinical skills.

To counter these threats to validity, two different kinds of evidence are sought to validate the construct: convergent evidence and discriminant evidence. This type of evidence was introduced through an empirical test from Campbell and Fiske (as cited in Angoff, 1988). They suggested that the correlations among different methods of measuring the same construct (convergent validity) would be higher than the correlations between different constructs measured by the same method. (discriminant validity). As Messick describes it, one is to "assess the degree to which the construct's implications are realized in empirical score relationships and the other to argue that these relationships are not attributable instead to distinct alternative constructs" (1989, p. 34).

Another way to assess for construct-irrelevant variance is through studies of differential item functioning (DIF). Construct-irrelevant variance can be investigated through isolation of a potential nuisance variable. For example, word problems on a math test may adversely affect the examinee with weaker reading comprehension skills and, therefore, not accurately represent that examinee's level of math ability. DIF analyses can assist in identifying these discrepancies.

Determining whether a test score interpretation is valid or not is a subjective process based on the amount and sources of validity evidence. It is, therefore, less appropriate to speak in terms of a score being valid or not valid and more appropriate to speak in terms of amount of evidence to support the likelihood of score validity. Score interpretation validity is reflected over a continuum and is not a discreet declaration.

The previous discussion presented the prevailing view of a unified concept of construct validity. It described the six sources of evidence of score validity, threats to validity, and ways to counter those threats. Attention is now turned to the second component of this study, the instrument to which this concept of validity will be applied. Following is a discussion of what is already known about reliability and validity issues surrounding OSCE score interpretation.

The OSCE: Reliability and Validity

The OSCE has been critical to the advancement of clinical skills assessment. It has proved extraordinarily useful since its introduction over 30 years ago, and its popularity continues to grow. It is, however, imperfect and continually evolving. Prior to its inception, there was no real means of measuring something other than knowledge that was fair and objective. Something was needed that would do for clinical skills assessment what the multiple-choice question did for knowledge assessment. The OSCE was a move towards this goal.

The OSCE

The OSCE was introduced into the medical literature in 1975. Harden et al. (1975) described the OSCE as a way to measure clinical competence that would assess a

wider range of knowledge and skills than traditional measures were able to do. The OSCE, as it is administered today, has remained fairly true to Harden's conception. Generally, students rotate through a series of stations, with each station representing a medical case to be evaluated and diagnosed. The number and types of stations vary depending on the program and the purpose. Some stations use a standardized patient – an actor playing the role of patient – while others may consist of a computer simulation or a written case history for the student to read and evaluate. The student's performance is generally evaluated by the standardized patient and/or an expert-evaluator, often a physician, using a checklist of behaviors that are evaluated as "did" or "did not" demonstrate. Some cases may require the students to write a summary of the case and then their findings and diagnoses. The write-up is then scored by an expert-evaluator. Where in the medical education program the student is required to complete the OSCE and whether passing the OSCE is required to advance to the next level of training varies from program to program depending on the context and purpose of the particular OSCE.

As increasing importance is placed on evaluating the qualities that cannot be captured through traditional test formats, the OSCE and its ability to predict future clinical performance have become areas of great interest. Over the last thirty years, a growing body of literature has addressed the reliability and validity of the OSCE as an assessment instrument for a wide variety of purposes in a wide variety of settings. Purposes include not only instruction and performance evaluation, but also serve as requirements for the advancement to a next level of training, passing a clerkship, or professional licensure. The OSCE is used not only in undergraduate medical education

and residencies, but also in many other health professions, such as dentistry, nursing, and physical therapy. Some OSCEs are held at the end of an academic year to assess learning over the course of that year, while others are administered at the end of a specific clinical rotation. Still others are designed to measure a specific quality such as communication, interpersonal skills, or English proficiency. Some use standardized patients as scorers, while others use expert-evaluators. A review of the literature suggests that reliability and validity studies are occurring in all these settings and under all these circumstances.

While the proliferation of OSCE programs existing globally speaks to the OSCE's worth, the lack of uniformity creates complications for researchers. Cerilli, Merrick, and Staren (2001) capture this sentiment well: "These examinations given at varied institutions and studying different groups with variable training also differ in content, difficulty, and structure. Comparison of different institutions' OSCEs is problematic because of these inherent differences in construction" (p. 325). It is in this spirit that the following review of the literature proceeds.

Reliability Issues Surrounding the OSCE

It is a generally accepted concept that score reliability is a necessary but not sufficient condition for score validity. With respect to the OSCE, although there are many factors that may create variance in scores, the research on reliability falls into two main areas: the instrument itself and the raters. Before turning to these factors, several sources of measurement error (Feldt & Brennan, 1989) that are viewed as threats to the reliability of the test score will be discussed.

Threats to reliability. The first threat to reliability is random variance within each individual student. It may occur systematically for a student that experiences test anxiety or randomly in the case of a random distraction on test day. A second source of measurement error is variance in the student's testing environment. Testing room temperature, lighting, and noise levels can be an issue for any exam, but performance assessments, such as the OSCE, will likely involve an additional layer of staging issues. The success of the OSCE is dependent on the training of standardized patients, faculty, and staff, the logistical coordination of rotating the students through the stations, and the management of a significantly longer testing period. All of these areas are likely to influence student performance.

Third, measurement error may be found in the variance in the instrument itself because "...only rarely does a set of exercises or a sample of ongoing behavior produce a precisely accurate representation of the total domain for any individual" (Feldt & Brennan, 1989). Chance would suggest that a test may contain items that expose one student's strengths and another's weaknesses. This potential for the test to misrepresent the student's true ability will be a source of inconsistency in measurements. Finally, measurement error may be found in the subjectivity of scoring. In multiple-choice style exams, this is of little concern. In performance based assessments, however, scoring becomes key. In performance based assessment, the reliability of the scores depends at least in part on the reliability of the raters. It is these final two sources of measurement error, the instrument and the rater, that are critical to the reliability of OSCE scores.

The assessment instrument. Research on the OSCE as a measurement instrument includes the following topics: the number of items on a checklist, how the OSCE can be used in conjunction with other scores for greater reliability, by whom items are best created, and the number of medical case stations that should be included in a given administration. It was thought that the checklist format would allow for improved inter-rater agreement, thereby addressing concerns about rater reliability (Newble, 2004). Although the checklist presents challenges of its own, the OSCE is now generally accepted as the tool for evaluating clinical skills proficiency (2004). It was perhaps hoped that the checklist would do for performance assessment what the multiple-choice question did for knowledge based assessment.

Classical test theory proposes that a longer test provides greater reliability. Research suggests this may not apply to OSCE checklists, at least from a rater reliability perspective. To explore how number the number of checklist items per medical case contributes to the variance in raters scores, Wilkinson, Frampton, Thompson-Fawcett and Egan (2003) looked at data from four consecutive years of OSCE data from undergraduate medical students. The number of checklist items per station increased over the four years as a move towards greater objectification; however, this increase in number of items correlated with neither the station inter-rater reliability nor the correlation between the station and the aggregate OSCE score. This would suggest that increasing the number of items does not necessarily improve reliability.

A greater number of items on the checklist may not improve reliability, but combining OSCE scores with other assessments may have an impact. Wass, McGibbon,

Van der Vleuten (2001) found that OSCE scores could be combined with written paper-and-pencil types, such as multiple-choice, extended matching, short answer, and essay. Carefully, weighting items on the various components increased the reliability of the exam overall. Similarly, Wilkinson and Frampton (2004) found that a combination of exam question types improved prediction of subsequent performance. The OSCE better predicted subsequent performance than the essay question, the modified essay question, or the multiple-choice questions; however, combining the assessments provided the strongest predictive value.

The items that comprise any given OSCE medical case checklist are a reflection of the creators of those items. One study found that independent experts can agree on the importance of items (Valentino, Donnelly, Sloan, Schwartz, & Haydon, 1998). This would suggest that an OSCE checklist is not likely to contain material that only a single individual deems important. However, this study also found that a group of faculty members is better able to establish important performance items than a single individual is, suggesting that the most important items are most likely to appear on the checklist through group effort. Wilkinson et al. (2003) identified examiner involvement in case preparation as the most important factor contributing to inter-rater reliability. This factor was regarded as having more impact on increased reliability than even their amount of experience as an OSCE evaluator.

Directly related to score reliability is the number of medical cases included in an OSCE administration. A definitive number was not revealed in a review of the literature; however, Carraccio & Englander (2000) state that a large number of stations is needed to

cover the necessary material and ensure reliability of the individual students' scores. Carraccio and Englander (2000) question whether fewer than 10 stations would be able to incorporate sufficient material to suggest score validity. Wessel, Williams, Finch, and Gemus (2003) found that eight stations had low internal consistency reliability and were unable to predict clinical performance. Their research suggested that 20 stations may be required to achieve an acceptable level of internal consistency. Harden (1975) in his original description of the structured clinical exam suggests 16 stations. Newble (2004) asserts that an OSCE would need to be in the four to eight hour range – a length too long to be practical for most test administrators – to see reasonable reliability. The USMLE Step 2 Clinical Skills (CS) component, the OSCE used for national board certification, has 12 stations (including any piloted stations not included in the students score) and lasts approximately eight hours (USMLE, 2007).

Traditionally, clinical skills exams were done at hospitals using real patients. However, the lack of standardization in patient encounters can lead to more unreliable examinee performances and ratings. Therefore, these days standardized patients are used for clinical examinations.

The evaluators. Traditionally, students were evaluated by one or more physician observers circulating through the hospital with the students. These days, due to the great demands placed on a physician's time, examiners are turning to the standardized patient (SP). The benefits of the SP over a real patient include increased inter-rater agreement and the reduced need for physician involvement. (McLaughlin, Gregor, Jones, & Coderre, 2006). The SP is often not only required to play the role of patient, but is also

expected to serve as an evaluator along with or instead of a physician. The primary disadvantage is that the SP may not be adequately trained or experienced to judge various levels of skills and/or provide feedback.

Studies have looked at the correlations between SP and physician raters' scores, the students' reactions to the role of the SP instead of a physician, the knowledge and expertise of the SP, and whether the SP is better qualified for some areas of assessment than for others. Research looking at reliability and the type of rater shows that there are differences in OSCE scores assigned.

Some studies have found only weak correlations between physician observer and SP scores (McLaughlin, Gregor, Jones, & Coderre, 2006; Thistlethwaite, 2002). Results from other studies challenge the notion that SPs are able to make reliable and valid judgments about an examinee's performance (J. A. Martin, Reznick, Rothman, Tamblyn, & Regehr, 1996). When comparing scores to a "gold standard," they found that physician observers were less likely to differ from the gold standard than were SPs. McLaughlin, Gregor, Jones, and Coderre (2006) found that SPs tended to rate students significantly higher than did physician-evaluators. They also found that physician scores were able to predict performance on a subsequent summative problem solving multiple-choice test, but SPs were not.

Rothman and Cusimano (2000) looked at physician scores and SP scores on OSCEs assessing international medical graduates' communication skills, broken down into two components: English proficiency and patient interview performance. The physician and SPs scores were highly correlated on the English proficiency but were

poorly correlated on patient interview performance. When English proficiency was explored more specifically, they found sizeable correlations between physician and SP scores but disagreement on the identification of problem students. SPs were more likely to rate examinees as problematic than were physicians. (Rothman & Cusimano, 2001).

While the research suggests that there are differences between SP and physician scores, there does not seem to be an examinee perception that the physician is preferable to the SP. McLaughlin et al. (2006) surveyed students at the end of a third year medical student internal medicine OSCE and received feedback on how the students felt about the SP examiner. They found that students did not find the station any more stressful than with a physician examiner. The majority of students felt that the SPs were as good as physician examiners and that they were sufficiently trained to judge. Lastly, a third of the students would have liked to have see more SP examiner stations.

McGraw and O'Connor's (1999) also explored examinee perceptions of SPs versus physicians. They performed a study that had a control group of students using real patients and an experimental group using SPs. Students were found to be no less satisfied with the SPs and in some areas had more positive experiences, as in the area of evaluator feedback. Furthermore, no significant difference was found between the OSCE scores of both groups.

Much of the literature focuses on the qualities, benefits, and disadvantages of the type of rater; however, other issues such as examiner fatigue were also explored. Humphris and Kaney (2001) looked at the time slots during which students rotated through a particular evaluator's station during the OSCE. They found no significant

relationship between the time slot and the score, suggesting that examiner fatigue was not an issue.

One of the ways to judge the training or quality of the raters is through rater agreement data. Inconsistencies between raters, or inconsistencies in a single rater's scoring over multiple performances, affect the reliability of scores themselves. Without inter-rater agreement data, how much of the variability in scores is attributed to actual student performance variability and how much is a result of rater inconsistency can not be determined.

Although examinees did not necessarily prefer physicians over SPs as raters, research suggests that the SPs provide less valid scores than do physicians. As we have seen, the research does not suggest a definitive conclusion regarding a preferred type of evaluator. Generally, the research does not suggest that the SP is a perfect substitute for the physician, nor does it suggest that the physician-evaluator is the only way to go. For now, programs will use both and continue to look for ways to improve rater reliability and validity. The literature does suggest that programs are looking seriously at SP evaluators as an alternative to physician testing and continue to provide training. Research has shown that training can be effective for reducing the variability in scoring, at least for trained staff, medical students, and lay persons, if not for physicians (van der Vleuten, van Luyk, van Ballegooijen, & Swanson, 1989). In any case, further study regarding the issue of physician-evaluator versus SP evaluator is warranted.

In summary, the OSCE reliability literature focuses on the characteristics of the instrument and the raters. It suggests that increasing the number of checklist items does

not necessarily increase reliability; that levels of item importance can be agreed upon by multiple independent experts; and that combining OSCE scores with scores from other assessments using multiple question types may be preferable to using the OSCE score alone. There does not seem to be a definitive number of cases established for assuming reasonable reliability; however, there is evidence that ten stations is an acceptable minimum. Regarding the raters themselves, research findings seem to support the inclusion of the SP in the OSCE, even though there are inconsistencies between the ratings by SPs and the ratings by physicians. Furthermore, students perceive the use of SPs positively.

Validity Issues Surrounding the OSCE

Of Messick's six sources of validity evidence, few are explored in the OSCE literature. It is generally accepted that most OSCEs are designed by physicians and experts in their fields, and that evidence of content validity can be presumed. That said, most studies are focused on external sources of construct validity and generally look to correlate the scores from an OSCE experience with scores on other assessments that ostensibly measure the same domain. Understandably, there is great interest in the ability of scores from the OSCE to predict future performances. Little research has been done, however, that explores the structural sources of construct validity of OSCE score interpretation. The following review of the literature is broken down into two parts. First, prior studies involving structural sources of score validity are reviewed. These studies look primarily at the issue of the checklist item as compared to the global item. Second, prior studies involving external evidence as a source of score validity are reviewed. These

studies tend to explore the predictive value of the OSCE in residency training, the predictive value in undergraduate medical education, and the ability to discriminate between levels of training.

Checklist item versus global rating. Some attention has been paid to the structural aspect of construct validity in that researchers are trying to refine the type of evaluation to use. Specifically, how does the checklist item compare to the global rating. With the checklist item, the student is expected to do a specific task or ask a specific question. The student is scored as "done" or "not done." The global rating is more likely to use a scale such as exceeds expectations/meets expectations/needs development. It is more likely used to assess a general dimension of performance, such as communication, empathy, or verbal expression (Hodges, Regehr, McNaughton, Tiberius, & Hanson, 1999). Although the checklist is widely accepted as the means of scoring the OSCE, there is concern that only criteria that are easy to define are included in the checklist and that detailed checklists, while reliable, may not truly reflect ability level (Newble, 2004). Specifically, several studies have suggested that a global rating may be better than a checklist for assessing qualities such as communication and interpersonal skills.

Some studies looked specifically at the issue of global scores and communication. Cohen, Colliver, Marcy, Fried, and Swartz (1996) looked at fourth year medical students and found that global ratings were better than checklists for evaluating interpersonal and communication skills. Mazor, Ockene, Rogers, Carlin and Quirk (2005) administered an OSCE to fourth year medical students, looking specifically at communication skills. They

explored whether the communication checklist items correlated with patients' perceptions of communication effectiveness and found that the two were not necessarily related.

Other studies looked at how global scores were able to distinguish between different levels of training and proficiency. Regehr, MacRae, Reznick, and Szalay (1998) looked at surgery residents' surgical skills and found that the global scores had equal or higher inter-station reliability and better predicted residents' training levels. Hodges and McIlroy (2003) found that the OSCE checklist items were not able to distinguish between third and fourth year medical student performance. They found, however, that the OSCE global scores were able to distinguish between the two levels of students on "coherence and questioning" but identified no difference between students on "empathy and non-verbal skills." This is perhaps not surprising, since empathy may not be a quality that increases incrementally with training and may be a more constant individual trait. Another study showed that global ratings are able to distinguish between interns (residents in their first year of post-graduate medical education) and family physicians but not between interns and residents (Hodges, Regehr, McNaughton, Tiberius, & Hanson, 1999). This same study suggested that checklists were relatively ineffective for evaluating residents and physicians because they scored lower than the interns. In other words, the score declined with as level of expertise increased. It was suggested that this is likely because experienced clinicians do not need to ask as many questions to get to their diagnoses. Hodges et al. determined, however, that it is likely that checklists were able to discriminate a more advanced intern from another, as more proficient interns obtained higher marks on the checklists than did their peers.

In summary, this review suggests that global scores may be the preferred method for assessing communication, personal interaction, and qualities such as "coherence and questioning." It also suggests that global ratings may be able to distinguish between disparate levels of training and that checklists may not be able to differentiate between those at a higher level of training. The research does not, however, suggest that global ratings are preferred for history taking and physical exam skills assessment at the medical student level.

The discussion turns now to literature that explores the external aspect of construct validity. Three broad areas seem to prevail: 1) prediction of post-graduate performance in residencies, 2) correlations with other test scores measuring similar domains, and 3) differentiating between levels of proficiency and training.

Predictive value in post-graduate medical education. The goal of the OSCE is not just have a student to do well on the OSCE, rather it is an attempt to create a "better" prepared/trained physician. One of the best criterion then, is not other scores obtained in the student setting, but rather, how students are doing after graduation. As such, some research explores the ability of OSCEs to predict performance at the residency level.

Taylor, Blue, Mainous, Geesey, and Basco Jr. (2005) looked at intern performance as judged by the residency director evaluation form and compared those to the scores from the prototype of the USMLE Step 2 CS component. They looked at the portion of the prototype that measured History and Physical Exam skills and used a checklist. Interpersonal skills were measured using two questionnaires. They found that the portions of the USMLE Step 2 CS prototype that used the checklist and measured

History and Physical Exam skills did not correlate well with other measures of future performance. The interpersonal scores (global scores) from the USMLE Step 2 CS prototype did, however, did appear to be a good predictor of resident director evaluations.

Another study found that OSCE scores correlated poorly with traditional final exams and correlated well with residency directors' feedback of first year postgraduate performance scores. This suggests that the OSCEs are assessing different clinical domains than traditional final exams (Probert, Cahill, McCann, & Ben-Shlomo, 2003).

Rifkin and Rifkin (2005) had internal medicine interns rotate through a four station OSCE. They then compared these scores to their USMLE Step 2 Clinical Knowledge (CK) scores taken the previous year. Not surprisingly the USMLE Step 2 CK score did not predict performance on the OSCE. This is representative of discriminant evidence; it is expected that a knowledge-based test is representing a different domain than a skills-based test.

Predictive value in undergraduate medical education. OSCEs are often administered upon completion of a particular year of study. Given at the end of a year of study, the OSCEs tend to be general and cross many clinical disciplines. Other OSCEs are administered as part of a required clerkship rotation. Given during clerkship rotations, the OSCE stations are specific to the particular specialty. Clinical clerkship rotations often have final exam grades and overall course grades. These assessments are often compared to OSCE scores.

One study specifically looked at the relationship between an OSCE given at the end of second year and exam scores from five required clerkship rotations. The findings

suggested that the general OSCE was able to identify students who would go on to fail one or more subsequent examinations in the required clerkships (I. G. Martin & Jolly, 2002). This is important because students that are identified early can be targeted for remediation. Walters, Osborn and Raven (2005) found that there was a moderate correlation between the score on the Psychiatry OSCE and the final grade for the Psychiatry clerkship.

Despite findings that suggest OSCE scores tend to correlate with other measures of clinical performance, the research does tend to be mixed. Wessel et al. (2003) found that OSCEs administered with first year physical therapy students did not predict subsequent clinical performance. And while the research generally suggests that the OSCEs do not have significant relationships with assessments of knowledge, again, there is research that suggests otherwise. Gerrow, Murphy, Boyd, and Scott (2003) found significant correlations between scores obtained from dental school OSCEs and the 300 question, multiple-choice exam administered by the National Dental Examining Board of Canada.

Discrimination between levels of proficiency. The OSCEs in many settings are criterion based. This means that students are expected to meet a particular standard to "pass." Many programs, including the USMLE Step 2 CS, assign simply a grade of pass or fail. There is research that suggests, however, that OSCEs are able to discriminate between levels of proficiency.

Joorabchi (1991) ran a 42 station pediatrics OSCE and found that the OSCE was able to distinguish between medical students and each of three classes of residents. This

was compared to another form of assessment, the resident performance ratings, which could only distinguish first year residents from third year residents.

The surgery specialty has produced research regarding the OSCEs' ability to discriminate between levels of residents. Sloan et al. (1996) found that a surgery OSCE was able to distinguish between students and different levels of residents. The performance scores for the four levels of trainees – junior medical students, first year residents, junior residents, and senior residents – improved significantly with level of training. Other research in the field of surgery demonstrated that the OSCE scores differentiated between students that received the lowest final grade in the surgery clerkship and those that received the highest (Merrick, Nowacek, Boyer, & Robertson, 2000).

In another study in the surgery specialty, Cerilli, Merrick and Staren (2001) correlated OSCE scores with the residents' number of years of training ranging from one year to five years. The OSCE consisted of both technical stations and clinical stations. The overall score correlated positively with year of residency. The technical stations, however, correlated significantly more with level of training than did the clinical stations. They also found that junior residents performed better on clinical stations than they did on technical stations; for senior residents it was the opposite. As discussed previously, this may be explained in that as level of expertise increases less of the checklist type items are needed for accurate assessment and diagnosis.

Despite the growing acceptance of Messick's unified approach to construct validity, surprisingly few studies gather evidence from the full range of sources of

evidence. This is perhaps because of the difficulties associated with seeking evidence of generalizability and with identifying consequences. Most studies, therefore, seem to focus on content and external sources, with limited studies of structural and substantive evidence. One study of the OSCE by Auewarakul, Downing, Jaturatamrong, and Praditsuwan (2005), however, very specifically approached the validity issues using Messick's view of validity as a unified concept. They sought evidence from as many of the sources as possible with the understanding that it was this collection of evidence that would support, or not support, any claims of validity. "It viewed validity as [a] hypothesis and sought specific sources of evidence to support or refute the proposed score interpretations" (p. 281). This study looked at several assessment types in evaluating third year medical students, including multiple-choice questions, essays, OSCEs and others. They then sought evidence of validity from multiple sources for each type of assessment. The results were mixed; however, the results may be less important than the approach used to explore the validity issue. Where many studies fall short, this study is an example of increasing confidence in score interpretation validity by supplying evidence from multiple sources and truly applying the principles of a unified concept of construct validity.

In summary, the literature suggests that the OSCE is able to predict some residency performance from undergraduate medical student assessments. The literature is mixed as to how the OSCE correlates with other assessments. Finally, the OSCE appears to distinguish between different levels of proficiency at some levels of expertise.

A review of the literature reveals a focus on the external aspect of construct validity, generally explored by the correlation of OSCE scores with other assessment scores. Less was found that considered structural or substantive evidence as a source of evidence of score validity. The following discussion describes how classical test theory principles can be used to analyze the assessment instrument for evidence of structural score validity. It will then describe how structural equation modeling techniques can be used to analyze the assessment instrument for evidence of substantive score validity.

Analysis: Classical Test Theory

Classical Test Theory has been the basis of test development for nearly 100 years. It is based on a set of assumptions upon which principles of test construction and score interpretation reliability and validity have been built (Allen & Yen, 2002). It is based on the premise that every observed test score is comprised of a "true" score plus some amount of error. More specifically, the observed score, X, is the actual score of an examinee on a given test. The true score, T, is the average of all observed scores and is a theoretical, not a practical, number. The error of the measurement, E, is the difference between the two. The following discussion will briefly describe each of these elements. It will then explore how classical test theory is applied to test construction, including a discussion of item difficulty, item discrimination, and internal-consistency reliability.

Classical Test Theory Assumptions

The following seven assumptions provide the foundation for test construction and test score reliability and validity in classical test theory (Allen & Yen, 2002). The first

two assumptions address the relationships between the true score and the error for one student on one testing occasion.

Assumption 1 states the relationship between the observed score, X; the true score, T; and the measurement error, E. The observed score is the sum of the true score and the measurement error. Assumption 2 provides the definition of the true score as the average of the scores that would be obtained if an examinee were to take the same test an infinite number of times. It is a theoretical, expected observed score, not one that could actually be obtained. It is important to note that the true score may or may not reflect the totality of knowledge or ability of the examinee. For example, the "ceiling effect" may have an influence if the test is too easy. In this case two examinees of unequal knowledge may actually obtain the same score as the test is unable to distinguish between the students with high scores.

Assumptions 2 through 5 define error of measurement as unsystematic and random variance that affects an observed score and prevents access to the true score. Where assumptions 1 and 2 address relationships between the true score and the measurement error for one student and one testing occasion, assumptions 3 through 5 address the relationships between the true score and the measurement error for many examinees on one testing occasion.

Assumption 3 states that there is zero correlation between true score and measurement error. In other words, the errors of measurement among examinees should be random, and not systematic. Assumption 4 states that there is zero correlation between the measurement error on test 1 and the measurement error on test 2; assumption 5 states

that there is zero correlation between the true score on test 1 and the error measurement on test 2.

Assumptions 6 and 7 define parallel tests and essentially tau (τ)-equivalent tests respectively. Tests are parallel if the true scores and the error variances are the same. Tests are essentially τ-equivalent if the true scores are the same except for an additive constant. Assumptions 6 and 7 assume two tests and are less applicable to this particular research study.

These seven assumptions are the foundation of classical test theory. These assumptions about observed scores, true scores, and measurement error and, as an extension the score variances, allow for a number of conclusions to be drawn. One of these conclusions suggests that reliability is viewed as the ratio of the variance of true scores to the variance of observed scores (Allen & Yen, 2002). This concept becomes important as the discussion turns next to test construction.

Classical Test Theory Test Construction

Reliability and validity in classical test theory rely on these assumptions and resulting conclusions about variability in test scores. Without variability in the test scores, there is no basis for analysis. Limitations for criterion referenced assessments such as the OSCE include the fact that most students meet the criterion and do well, creating a limited range of scores and minimal variance. However, even criterion-referenced assessments are developed using classical test theory processes. Therefore, when scores have been generated from a test administration, reliability is analyzed and an item analysis is performed.

Reliability. These analyses provide two sources of evidence for the validity of scores. As mentioned in the previous section, reliability can be viewed as the ratio of true score variance to observed score variance; however, because true scores, and hence error variances, exist in theory only, estimates of reliability must be generated instead. Internal-consistency reliability known as Cronbach's coefficient alpha (α) is commonly used as an estimate of reliability. The formula for calculating coefficient α takes into account the variance of the individual items and the variance of the test scores overall to estimate the test's internal consistency. Coefficient α provides good estimates of reliability when the items intercorrelate highly; in other words, the test content is homogeneous and the test is measuring a single dimension. In medical education, students are from a homogeneous group and are generally quite able. This restriction in range of ability, and therefore restriction in variance of scores, may negatively influence the accuracy of the resulting reliability estimate derived from the alpha coefficient. Reliability is used in conjunction with item difficulty and item discrimination values to determine whether items are contributing value to the measurement instrument.

Raykov (2004) suggests that the alpha coefficient is not the most effective statistic for assessing internal consistency. Many tests violate the item homogeneity assumption; therefore, Raykov proposes the use of maximal reliability estimates using weighted item reliability estimates. However, when weighted maximal reliability estimates are generated, the same item weighting must be used to generate scores from tests. Although theoretically interesting, the formative use of OSCE scores may not warrant a sophisticated scoring model.

Item analysis in classical test theory, an item analysis is conducted using the item scores from a test administration. The primary analyses performed in an item analysis are item difficulty and item discrimination. Item difficulty analyses help determine whether items are in an appropriate range for the construct being measured. Discriminant analyses indicate how well each item contributes to the overall test score.

Item difficulty. Item difficulty is described as the proportion (p_i) of examinees that provide a correct response to an item or earn the highest possible score; the higher the proportion, the easier the item. Items with a p_i of 0.0 or 1.0 should be considered for modification or removal from the test, as no information about the differing abilities of the examinees is provided. With a typical norm referenced test and normally distributed examinee pool, a p_i between 0.30 and 0.70 with an average of about 0.50 is considered ideal, as it will maximize the information provided about the ability of the examinees (Allen & Yen, 2002). A criterion referenced test and highly able examinee pool will likely generate higher item means. In a criterion based test with a pool of highly able students, a moderate p value may be seen as falling between 0.60 and 0.95. This suggests that a reasonable number of students is getting the item correct..

Item discrimination. The ability of an item to discriminate between levels of examinee knowledge or skill is generally assessed using item/total-test-score correlations (r_{iX}). The expectation is that examinees of higher ability will be more likely to answer an item correctly than examinees of lower ability. When this is the case, the r_{iX} is positive. The higher the item discrimination value, the greater the ability of the item to discriminate between levels of ability. A high item discrimination value suggests that the

item is measuring what it is supposed to be measuring and is a source of evidence for score validity. A negative item discrimination value indicates that the lower scoring students are getting the item correct while the higher scoring students are getting it incorrect. A negative correlation may indicate that an error in scoring was made or that some factor besides examinee ability on the relevant construct is being captured. An item/total-test-score correlation of zero suggests that there is no relationship between the item scores and the test scores.

With item discrimination, a high or low item *p* value with a corresponding low item/total-test-score correlation is considered normal. When only a small percentage of examinees get an item correct or incorrect, it does not matter whether they are from a group with an overall high or an overall low score, the correlation is likely to be low. More relevant to identifying potentially problematic items is the occurrence of items with a moderate *p* value and a corresponding low item/total-test-score correlation.

In typical test construction, item difficulty, item discrimination, and reliability are used together to determine an item's contribution to the measurement scale. An item should be considered for modification or removal from the test if the item difficulty is moderate *and* the item discrimination value is low *and* if the reliability of the test would increase if the item were removed.

Prior research that has explored the internal structure of the OSCE has looked only at the medical case level. For example, Auewarakul, Downing, Praditsuwan, and Jaturatamrong (2005) used the principles of classical test theory and item analysis to improve the reliability of an internal medicine undergraduate OSCE. They performed an

item analysis, treated each of the 25 OSCE stations as an item, and were able to eliminate problem cases and increase overall reliability. A review of the literature did not reveal a study that looked at the OSCE at the checklist item level.

Classical test theory and external evidence of validity. Evidence of construct validity in classical test theory is often sought through external sources. Test scores are correlated with scores from other tests that ostensibly measure the same construct. High correlations suggest the ability to predict one score from the other. They also suggest that the score interpretations will be similar.

Analysis: Confirmatory Factor Analysis

Structural equation modeling (SEM) refers to a family of related procedures rather than to a single statistical technique (Kline, 2005). This family of procedures is broad in scope and includes the techniques of the general linear model (e.g. regression analyses, multivariate analysis of variance, and exploratory factor analysis) as well as causal models (e.g. path analysis), and confirmatory factor analysis (CFA). In SEM, samples sizes exceeding 200 cases are considered large; samples of below 100 cases are considered small (Kline, 2005).

CFA is a technique generally used in the advanced stages of the research process to test a preconceived theory. Relationships between the variables and the factors are presupposed, and hypothesized models of the data are presumed, based on theory and/or empirical data. In CFA, a factor is likely a "latent variable," a trait or quality that influences the responses to items on an assessment. It cannot be observed directly, but can be inferred through observed variables (Kline, 2005).

Assigning a single score to an assessment that measures multiple dimensions is unlikely to allow for a valid test score interpretation. CFA serves a purpose similar to a classical test theory based item analysis when used in test construction or scale development (Stevens, 2002). Theory dictates which items or subscales are likely to group, or "load," together onto a particular subscale or "factor." Statistics are generated to support or refute that theory.

Models are tested through CFA by generating two or more models for the data. Certain components of an assessment may correlate, or load, more highly on one factor than on others. CFA allows for the comparison of two or more competing models. For example, it may compare one model specifying that subscales load on a single dimension, or, factor, and a second model specifying that the subscales load on multiple factors based on theoretical foundations.

A covariance or correlation matrix serves as input into the program along with the standard deviations of each item on the assessment. The independent variables (the subscales) and dependent variables (the factors on which the items are to "load") or, latent variables, are specified. Model fit to the data is assessed using several statistics. Comparing the fit statistics of the competing models provides a quantitative means of assessing the data that should be used along with sound theoretical reasoning.

The CFA generates fit statistics to evaluate how well the theoretical model fits the data. Some fit statistics apply to the individual model parameters, while others apply to the overall model. With respect to individual model parameters, the factor loading tells how much of the variance in the observed subscale score is explained by the model. The

factor loading value is the correlation between the observed variable (the subscale score) and the factor, and should be relatively high. Relatively high factor loadings between the subscale scores and the factor indicate the subscales are measuring a common underlying factor and suggest evidence of convergent score validity. In addition, the correlation between the two factors should be less than 0.85 which provides evidence of discriminant score validity as the factors would suggest measuring different dimensions.

Measurement error variance describes the error associated with each variable. It describes the part of each observed variable not explained by the factors and may also include error from the unreliability of the observed variables.

With respect to fit statistics that apply to the overall model fit, the chi-square (χ^2) statistic tests the hypothesis that the proposed model fits the pattern of the covariation in the observed scores. Therefore, a low χ^2 value is desired, which would indicate that the data are not different from the proposed model. In this case, a null hypothesis suggesting that the model fits the data would not be rejected.

The root mean square error of approximation (RMSEA) estimates the lack of fit between the hypothesized model and the observed data. Browne and Cudeck (as cited in Kline, 2005) suggested that an RMSEA ≤ 0.05 indicates approximate fit and an RMSEA ≥ 0.10 suggests poor fit.

The comparative fit index (CFI) estimates the improved fit of the hypothesized model over a null model where all the variables are assumed to be uncorrelated. A higher value indicates more improvement in model fit of the hypothesized model. In general, a CFI ≥ 0.90 suggests relatively good fit (Stevens, 2002).

McKinley and Boulet (2005) used CFA techniques to improve the psychometric qualities of OSCE scores. They were able to identify the multidimensionality of the OSCE cases, and thereby provide additional evidence for construct validity.

This chapter provided a review of what is already known regarding the validity and reliability of OSCE scores and the role of classical test theory and confirmatory factor analysis in providing validity evidence. This review of the literature also suggests that several questions relevant to the reliability and validity of OSCE scores remain: What is the internal consistency of checklist scores on different OSCE stations? In other words, does the OSCE measure a unitary dimension or are there multiple dimensions of performance inherent in the OSCE? Are all items on the OSCE checklist for a standardized patient encounter equally predictive of final judgment about student proficiency? How well do OSCE scores from a refined checklist predict other judgments of student proficiency? It is these questions that are investigated in this study.

Chapter 3: Methods

The data for this study came from Objective Structured Clinical Examination (OSCE) administrations at the University of Washington. The use of the data for this purpose was approved by the University's Human Subjects Division. Following are descriptions of the subjects, the specifics of the OSCE medical cases, and the analyses used in the study.

Subjects

The subject pool was comprised of two cohorts of medical students from the University of Washington, School of Medicine. The vast majority of students were residents of a five-state region that included Alaska, Idaho, Montana, Washington, and Wyoming. The students participated in the Objective Structured Clinical Exam (OSCE) at the beginning of their senior year of medical school in the 2005-06 and 2006-07 academic years. The majority of each cohort entered medical school in 2002 and 2003 respectively. One hundred eighty one students participated in the 2005 OSCE administration and 165 students participated in the 2006 OSCE administration. Thus, scores for 346 students were available for this study. Of the pool of students, 162, or 46.8 percent were male. The average examinee age at the time of the OSCE was 28, with ages ranging from 23 to 48. The median age was 27 and the mode was 26.

OSCE Medical Case Descriptions

The scores generated from the 2005 and 2006 senior medical student OSCE administrations were obtained. The 2005 and 2006 OSCE administrations required students to rotate through six medical case stations; there were four medical cases in

common between the two years. It is the data from the four common cases that were used in this research project. OSCE items for common cases were included in the analyses only if scores were available for both the 2005 and 2006 cohorts. The scores for the two cohorts were aggregated and treated as one sample, as prior OSCE studies have done (Park et al., 2004), in order to enhance the power of the statistical results.

Each student was assigned an "overall rating" for each medical case. The overall rating is a holistic evaluation assigned to each student based on his or her performance on a particular medical case. Two of four common cases (cases A and B) required both a standardized patient (SP) encounter and a write-up. A write-up required that the student, upon completion of the SP encounter (if applicable), write up the case, including a history, a physical exam (if applicable), a diagnosis, and a follow-up treatment plan. In case A, the SP encounter and the write-up were scored by physician-evaluators using checklists. The physician-evaluators used both the SP encounter checklist and the write-up checklist scores to assign the overall rating. In case B, the SP encounter was scored by a physician-evaluator. Each write-up was scored by two physician-evaluators; the scores were added together for this analysis. Again, the SP encounter checklist and the write-up checklist scores were taken into account for the overall rating.

One of the four cases (case C) required only an SP encounter. In this case, the physician-evaluator scored the SP encounter using a checklist and then assigned the overall rating based on the SP encounter. One of the four cases (case D) required only a write-up. In this case, the physician-evaluator scored the write-up using a checklist and assigned the overall rating. All four cases had a History component. Cases A and C

included a Physical Exam component. (Case D also had a Physical Exam component but it was not reflected in the write-up checklist.) Cases A, B, and C each had two Introductory items and two Impressions items (communication and organization).

Some items on the SP checklist and write-up evaluation checklist were dichotomous (done/not done); others were on a rating scale (exceeds expectations/meets expectations/needs development). Items that were done/not done were scored 1 or 0 respectively. These types of items were assessing whether the student performed a particular task or asked a particular question of the standardized patient. Items scaled exceeds expectations/meets expectations/needs development were scored as 2, 1, or 0 respectively. These types of items were more likely assessing a characteristic such as communication skills or organizational flow.

Items in the Introductory, History, and Impressions components of Cases A, B, and C were identical, except for Item 10 which did not appear on Case C. These items were scored as done/not done, except for the items in the Impressions component which were scored as exceeds expectations/meets expectations/needs development. A Physical Exam component was included in Cases A and C. The checklist items included in that component were unique to each case. The write-up section of Cases A and C had four out of five items in common. The write-up section that comprised Case D in its entirety was unique to that case. Table 1 summarizes the number of items and item type by case.

Table 1: Summary of Item Number and Type by Medical Case

	Case A		Case B		Case C		Case D	
	# Items	Type	# Items	Type	# Items	Type	# Items	Type
SP Checklist:								
Introductory	2	D/ND	2	D/ND	2	D/ND	-	-
History	13	D/ND	14	D/ND	12	D/ND	-	-
Physical Exam	8	D/ND	-	-	10	D/ND	-	-
Impressions	2	E/M/N	2	E/M/N	2	E/M/N	-	-
Overall rating	-	-	-	-	1	E/M/N	-	-
Write-up:								
Legibility	1	E/M/N	1	E/M/N	-	-	-	-
History	1	E/M/N	1	E/M/N	-	-	4	D/ND
Physical Exam	1	E/M/N	-	-	-	-	-	-
Diagnosis	1	E/M/N	1	E/M/N	-	-	3	E/M/N
Follow-up Plan	1	E/M/N	1	E/M/N	-	-	-	-
Ethics	-	-	-	-	-	-	4	D/ND
Overall rating	1	E/M/N	1	E/M/N	-	-	1	E/M/N

D/ND = done/not done
E/M/N = exceeds expectations/meets expectations/needs development

Analyses

Several sources of evidence of OSCE score validity were explored using the principles of classical test theory: content; substantive, using confirmatory factor analysis; and structural and external. Missing data were accounted for by pairwise deletion. Each of the four medical cases was explored independently of the others.

Structural Sources of Construct Validity

Principles of classical test theory were applied to explore the instrument at the item level. SPSS 11.5 was used for the classical test theory item analyses. By examining the internal structure of the instrument, problematic items were identified. Items that do not add value to the scale or items that detract from the reliability of the scale were flagged for removal from the analyses. All items were initially included in the reliability

and item analysis calculations to establish a base value of estimated reliability upon which to improve.

Reliability. Cronbach's coefficient alpha (α) was generated for each case to explore the internal-consistency reliability of the OSCE cases. In general, confidence in the reliability of the case is warranted at an alpha of 0.70 or greater. The alpha coefficient for each medical case was used in conjunction with item difficulty and item discrimination to determine whether an item should remain part of the instrument scale.

Item difficulty. Item means (p_i or p values) were generated for each of the items. A p value of 1.0 indicates an item that all students answered correctly and, therefore, the item has no variance. From a psychometric standpoint, items with a p value of 1.0 add no value to the measurement instrument and were flagged for removal from the analyses. Items with p values greater than 0.50 but less than or equal to 0.75 were considered items of moderate difficulty. Items with moderate p values were flagged for possible removal from the analyses if paired with a low item discrimination value.

Seventeen items common to three out of four medical cases were identified. Patterns in the p values of these common items across medical cases were noted. In addition, there were patterns of means within each medical case that indicated that possibly more than one dimension was being measured. These items were flagged for further analysis.

Item discrimination. An item/total-test-score correlation (r_{iX}) was generated for each of the items to determine how well each item was able to discriminate between levels of student ability. Items with correlations of less than 0.20 were considered low

and unable to effectively discriminate between levels of examinee ability. These items were flagged to be explored in conjunction with the item difficulty and case reliability to determine the item's contribution to the scale.

Alpha-if-item-deleted. This calculation was generated to identify whether the reliability of the scale would increase or decrease if the item was eliminated from the analysis. If the reliability would increase with the removal of the item, the item is not positively contributing to the scale. These items are adding no value and are, in fact, detracting from score reliability. The alpha-if-item-deleted value was used in conjunction with item difficulty and item discrimination values to determine whether the item should be removed from the assessment.

Process for item removal and subscale creation. Each of the medical cases was explored independently of the others. The item difficulty, item discrimination, reliability coefficient, and alpha-if-item-deleted values were considered together when refining the scales of the medical cases. Patterns were sought in the item difficulty, item discrimination and alpha-if-item-deleted statistics for possible subscales. Item groupings, the subscales, were identified using the item statistics as well as the content of each item. Item groupings had to be reasonable and suggest evidence of face validity.

Items that met predetermined criteria were considered problematic and were removed from the next analyses; therefore, they would not be included in a final scale or subscale. The criteria for item removal included a moderate mean ($0.50 < p_i <= 0.75$), a low item/total-test-score correlation ($r_{iX} <= 0.20$), and an alpha-if-item-deleted showing the item had a negative contribution to the overall case reliability. Once problematic

items were removed, a new scale and a new estimated reliability coefficient were generated. Again, any items identified as having moderate means, low item/total-test-score correlations, and an increased alpha-if-item-deleted value were removed. This process was repeated until no items met the criteria for removal.

Correlations between item and overall rating. Pearson product-moment correlations (r_{xy}) were run between each item and the overall rating. Since the overall rating was based on the items in that case, a high correlation would be expected. Results were explored to see how each of the items contributed to the overall rating. Existing patterns among correlations were noted.

Substantive Sources of Construct Validity

Substantive sources of construct validity were explored through confirmatory factor analysis. LISREL 8.3 statistical software was used for the confirmatory factor analysis. Using classical test theory principles, each OSCE case was refined and broken into subscales as the analyses suggested different dimensions. A confirmatory factor analysis was performed on the refined subscales to explore whether they suggested a single-dimension model or a multidimensional model. It was expected that a model suggesting that the revised subscales represented multiple dimensions would better fit the data.

PRELIS was used to prepare the covariance matrix for use by the LISREL software. The hypothesized models were determined to be overidentified. Based on theory and the empirical data from the structural analysis of the OSCE, certain hypotheses were generated about the underlying factor structure of the data. It was

hypothesized that the OSCE checklist was measuring different dimensions that could be described by the latent variables History and Physical Exam. It was also hypothesized which subscale scores would load onto which of the two latent variables. To see if this was indeed the case two competing models were hypothesized: a one-dimensional model and a two-dimensional model. The latent variables were free to correlate with one another, and the subscales were limited to load only on one factor.

Statistics with respect to the individual model parameters were generated and analyzed: factor loadings and measurement error variance. A *t* value of greater than the absolute value of 2 indicated whether the parameter was significant. Also reviewed was the correlation between the two hypothesized factors.

Overall model fit statistics were also explored: chi-square, root mean square error of approximation, and comparative fit index. These values were generated for each model and the two sets of results were compared for evidence of a model that best fits the data.

External Sources of Construct Validity

External evidence of score validity was sought through correlations with other medical school assessments. SPSS version 11.5 was used to explore the correlations of the refined medical case subscales with scores from the USMLE Step 2 Clinical Knowledge (CK) component. A numerical score was obtained and used for each USMLE Step 2 CK score. The Clinical Skills (CS) component of the USMLE Step 2 is most comparable to the OSCE. Correlations between USMLE Step 2 CS and OSCE scores would have provided convergent evidence of score validity; however, the lack of variance in the USMLE Step 2 CS scores prohibited that analysis.

In summary, all items from each medical case were initially included in a case analysis to establish a base value reliability estimate from which to start. Item difficulty and item discrimination values were considered along with the reliability of the medical case. Items were eliminated and/or grouped into subscales if the analysis supported that action. Revised scale analyses were rerun to generate increased overall medical case reliability. A confirmatory factor analysis was run on the revised subscales to explore whether the subscales represented a single or multiple dimensions. Scores were compared with other medical student assessments as a source of evidence of external OSCE score validity. Table 2 summarizes the steps taken to explore the reliability and validity of the OSCE scores.

Table 2: Summary of Steps Taken

Steps of Analyses
Step 1: Compiled 2005 and 2006 data
Step 2: Identified medical cases A, B, C, and D as common to both years
Step 3: Identified items for which there were data for both cohorts of students
Step 4: Ran item statistics using SPSS 11.5
Means Item/total-test-score correlations Alpha for the case overall Alpha-if-item-deleted
Step 5: Identified patterns for possible multidimensionality
Step 6: Identified problematic items: Moderate mean, $0.50 < p_i <= 0.75$ Low item/total-test-score correlations, $r_{iX} <= 0.20$ Alpha-if-item-deleted, increases
Step 7: Removed problematic items; reran data; repeated as necessary
Step 8: Ran correlations between items and medical case "overall" rating
Step 9: Ran confirmatory factor analysis on revised subscales
Step 10: Correlated revised subscales with USMLE Step 2 CK scores

Chapter 4: Results

The results of the analyses of the Objective Structured Clinical Examination (OSCE) were explored for evidence of score reliability and internal and external validity.

Results from Analyses

Each medical case of the OSCE was analyzed independently of the others. Means for items that were common across cases were examined. Means, item/total-test-score correlations, alpha coefficients, and alpha-if-item-deleted values were generated for each item to explore item contribution to the scale. As decisions were made to adjust the scales by creating subscales or eliminating problematic items, alpha coefficients were computed to examine the impact of scale adjustments on overall reliability. The correlations between each item and the "overall rating," a holistic evaluation of the examinee's overall performance by a physician-evaluator, were explored. A confirmatory factor analysis was conducted to investigate the validity of subscale scores. The scores from the revised scales and subscales were also correlated with USMLE Step 2 Clinical Knowledge (CK) scores as a source of external discriminatory evidence of score validity.

Internal Structure Analyses

The internal structure of the OSCE was explored at the item level. The criteria for item removal included a moderate mean ($0.50 < p_i <= 0.75$), a low item/total-test-score correlation ($r_{iX} <= 0.20$), and an alpha-if-item-deleted showing the item made a negative contribution to overall case reliability. The removal and grouping of items resulted in revised scales and subscales with signs of increased reliability. The internal consistency of the case scores suggested greater validity of OSCE case score interpretations.

Reliability coefficients. A reliability coefficient was generated for each case using all items. The reliability coefficients were $\alpha = 0.5333$ for Case A, $\alpha = 0.6924$ for Case B, $\alpha = 0.6173$ for Case C, and $\alpha = 0.3193$ for Case D. These were considered baseline values upon which improvement in scale reliability was to occur.

Item difficulty. The grand means of the item means for each medical case were 0.6895, 0.6769, 0.6701, and 0.6497 on Cases A, B, C, and D respectively. The grand means for each case are substantially higher than is typical in a norm-referenced test; however, these high means are not surprising, taking into account the criterion-based nature of the OSCE and the homogeneous, highly able examinee pool. Appendices A, B, C, and D show the means for all items on Cases A, B, C and D respectively.

Items 1 through 15 (dichotomous) and the two Impressions (polytomous) items on Cases A, B, and C were identical, except for item 10 which did not appear on Case C. Items 1 through 15 included Introductory questions, and History questions; the Impressions items asked for overall impressions of organization/flow and communication skills. Table 3 shows the means of each of the common items on Cases A, B, and C. Many items had similar means across items. For example, item 1 had values of 0.9653, 0.9855, and 0.9740 on Cases A, B, and C respectively. At the other extreme were items with a wide range of means: item 13 had values of 0.0665, 0.6821, and 0.5694 on Cases A, B, and C respectively. Items that show the widest spread of means over the three cases are bolded in Table 3.

Table 3: Means for Items Common to Cases A, B, and C

Item	Means			Spread of Means
	Case A	Case B	Case C	
Introductory	0.9653	0.9855	0.9740	0.0202
Introductory	0.9855	0.8728	0.9566	0.1127
History	0.9769	1.0000	0.9769	0.0231
History	0.6387	0.5260	0.4913	0.1474
History	0.9827	0.9682	0.9249	0.0578
History	0.9884	0.8179	0.8410	0.1705
History	0.9711	0.8960	0.8757	0.0954
History	0.6156	0.6012	0.5318	0.0838
History*	**0.8526**	**0.5462**	**0.4884**	**0.3642**
History	0.8064	0.6618	-	0.1446
History*	**0.9393**	**0.5000**	**0.7486**	**0.4393**
History	0.6705	0.4884	0.6416	0.1821
History*	**0.0665**	**0.6821**	**0.5694**	**0.6156**
History	0.7457	0.9191	0.9653	0.2196
History	0.6185	0.6329	0.6763	0.0578
Impressions 1	1.0405	1.0173	1.0780	0.0607
Impressions 2	1.2052	1.0954	1.2428	0.1474
Grand Mean	0.7405	0.6953	0.7213	

*Items with the greatest spread between highest and lowest means across the cases

Grand means for the medical cases and ranges of item means across similar items were noted. In addition, moderate item means were flagged to be explored in conjunction with item discrimination values and alpha-if-item-deleted values.

Item discrimination: The item/total-test score correlations for Case A ranged from r_{iX} = -0.1089 to r_{iX} = 0.4419 with a mean of 0.1582, for Case B r_{iX} ranged from of r_{iX} = 0.0832 to r_{iX} = 0.5310 with a mean of 0.2899. For Case C, r_{iX} ranged from r_{iX} = -0.0067 to r_{iX} = 0.5058 with a mean of 0.1837 and, for Case D, r_{iX} ranged from r_{iX} = 0.0340 to r_{iX} = 0.2335 with a mean of 0.1184. Appendices A, B, C, and D show the item/total-test-score correlation results for all items on Cases A, B, C and D respectively.

After evaluating dimensionality and reliability, the item/total-test-score correlation ranges and means of *revised* scales and subscales were generally higher for each medical case: Case A ranged from r_{iX}= 0.0701 to r_{iX} = 0.5704 with a mean of 0.3108; case B ranged from r_{iX} = 0.0716 to r_{iX} = 0.5494 with a mean of 0.3528; case C ranged from r_{iX} = 0.0516 to r_{iX} = 0.5181 with a mean of 0.2753; and case D ranged from r_{iX} = 0.0575 to r_{iX} = 0.3214 with a mean of 0.1842. All medical cases except Case B showed improved item discrimination values. Table 4 summarizes the item discrimination values before and after the revised scales were created. The mean item discrimination values of all the medical cases were higher in the revised scales and subscales.

Table 4: Summary of Item Discrimination Values for Original and Revised Scales

		CASE A	CASE B	CASE C	CASE D
Original	**Low**	-0.1089	0.0832	-0.0067	0.0340
	High	0.4419	0.5310	0.5058	0.2335
	Mean	0.1582	0.2899	0.1837	0.1184
Revised	**Low**	0.0701	0.0716	0.0516	0.0575
	High	0.5704	0.5494	0.5181	0.3214
	Mean	0.3108	0.3528	0.2753	0.1842

Alpha-if-item-deleted: Figures showing the change in case score reliability if the item were deleted from the scale were also generated. The initial run on Case A revealed eight items that were detracting from case score reliability. Cases B and C contained five such items, and Case D contained two. Appendices A, B, C, and D show the alpha-if-item-deleted results for all items on Cases A, B, C and D respectively.

Item difficulties were used in conjunction with item discrimination and estimated reliability values to identify potentially problematic items and to create potentially more

reliable and valid scales. When an item met the criteria for removal from the scale, the item was removed and the scale analysis was rerun. The following describes each case and how the items were removed and the new subscales generated.

Case A. The initial analysis included all 30 items from both the standardized patient (SP) encounter and the write-up sections of the medical case, to establish a base estimate of reliability upon which to improve. The reliability estimate (alpha) for Case A was $\alpha = 0.5333$. It was noted that the items from the case write-up were grouped together with four out of five items having low item/total-test-score correlations. Three of those four had moderate *p* values and an alpha-if-item-deleted showing an increase. The write-up items were, therefore, removed from the scale to be analyzed as a separate group of items potentially measuring a different dimension of performance than the SP encounter checklist measured.

The alpha coefficient for the remaining items was rerun. The Case A alpha coefficient improved to $\alpha = 0.5831$. However, the scale for this case showed a pattern of means suggesting that multiple constructs are being measured. To explore this possibility, reliability analyses were run after separating these items into subscales.

The first group of items included the two Introductory items and the 13 History items. The second group of items included the eight Physical Exam items and the two Impressions items. Reliability of the subscales dropped to 0.5686 and 0.5026 respectively. Removing the two Introductory items from the first scale and the two Impressions items from the second scale did not improve the reliability of either scale. Likewise, combining the two Introductory items with the second scale and moving the

two Impressions items to the first scale caused the reliability of both scales to drop dramatically.

The first scale of items 1 through 15 was divided into two sub-scales. Items 1 through 7 were run as one scale for a reliability of 0.4743. Item 4 was identified as having a moderate *p* value, a low item/total-test-score correlation, and an alpha-if-item-deleted showing an increase in scale reliability if the item was removed. Item 4 was, therefore, removed and the remaining items were rerun, generating a reliability coefficient of $\alpha = 0.7175$. These items grouped well and showed signs of measuring the same domain. The remaining History items, 8 through 15, were run producing a reliability estimate of $\alpha = 0.5446$. No items met the criteria for removal. These items comprised the second subscale.

A scale including the Physical Exam items, 16 through 23, and the Impressions items, 24 and 25, produced a reliability estimate of $\alpha = 0.5026$. Item 22 was observed to have a moderate *p* value, a low item/total-test-score correlation, and an alpha-if-item-deleted indicating an increase in scale reliability. Item 22 was removed from the analysis and the scale was rerun for an improved reliability estimate of $\alpha = 0.5114$. Item 19 now met the criteria for removal. Item 19 was removed and the scale rerun for an improved reliability of $\alpha = 0.5123$. No further items were identified for removal. These remaining items comprised the third subscale.

The write-up items were analyzed. In three out of the five items, *p* values were moderate and item/total-test-score correlations were low. Of those three, one also showed an alpha-if-item-deleted that indicated that reliability would increase. Overall reliability

was extremely low at $\alpha = 0.1156$. These items were, therefore, considered separately and not as a single subscale.

Table 5 presents the steps and decisions described above. The proposed structure for Case A has three subscales that are better able to reliably measure the ability of examinees on the skills measured by the OSCE. Items 4, 19, and 22 were removed from the analysis and the write-up items should be treated separately. Two of the three subscales showed improved reliability over the original unitary scale. The original scale for Case A had eight items with moderate *p* values and low item/total-test-score correlations. The resulting three subscales have only one between them.

Table 5: Summary of Iterations for Case A Estimated Reliability Improvement

Iteration	Items Included	α	No. of Problematic Items[1]	No. of Items for Removal[2]
1	All items 1-30 – Introductory, History, Physical Exam, Impressions, Write-up	0.5333	8	4
2	Introductory, History, Physical Exam, Impressions	0.5831	5	3
3	Introductory, History (items 1-15)	0.5686	2	2
4	Introductory, History (items 1-7)	0.4743	1	1
5	Introductory, History (items 1-7, item 4 removed)	**0.7175**	0	0
6	History (items 8-15)	**0.5446**	1	0
7	Physical Exam, Impressions	0.5026	3	1
8	Physical Exam (items 16-23, item 22 removed), Impressions	0.5114	1	1
9	Physical Exam (items 16-23, items 19 and 22 removed), Impressions	**0.5123**	0	0
10	Write-up (items treated individually)	0.1156	3	1

[1]moderate mean and low item/total-test-score correlation
[2]moderate mean, low item/total-test-score correlation, and alpha-if-item-deleted increases
Bold alpha coefficients indicate final revised subscale alpha coefficients.

Case B. The initial analysis included all items from both the SP encounter and the write-up to establish a base value reliability estimate upon which to improve. The reliability estimate for Case B was $\alpha = 0.6924$. As with Case A, the results of the reliability analysis suggested that the four write-up items were best removed and treated individually.

The four write-up items were removed. Additionally, item 3 was removed because all students demonstrated the behavior and, therefore, the item did not add to the reliability of the scale from a psychometric standpoint. The reliability analysis was rerun with these items removed. The overall estimate of reliability for Case B improved to $\alpha = 0.7639$. Item 15, however, now met the criteria for removal: a moderate *p* value, a low item/total-test-score correlation and an alpha-if-item-deleted indicating that overall reliability of the case would improve. Item 15 was removed from the analysis and the reliability estimate for Case B increased to $\alpha = 0.7695$.

An analysis was run on the four write-up items for a reliability estimate of $\alpha = 0.3952$. No items fit the criteria for removal. Given this low estimate of reliability, these items were to be treated separately.

Table 6 shows the steps described above for Case B. Case B did not include Physical Exam items. The History items do not show the same pattern as Case A where it appeared that the History items were measuring two different dimensions. The proposed structure for Case B with improved reliability estimates shows that items 3 and 15 should be removed, and the write-up items should be treated individually. The original scale had

two items with both moderate *p* values and low item/total-test-score correlations. The revised subscales have only one between them.

Table 6: Summary of Iterations for Case B Estimated Reliability Improvement

Iteration	Items Included	Alpha Coeff.	# of Problematic Items[1]	# of Items for Removal[2]
1	All items 1-22 – Introductory, History, Impressions, Write-up	0.6924	2	1
2	Introductory, History (items 1-16, item 3 removed), Impressions	0.7639	1	1
3	Introductory, History (items 1-16, items 3 and 15 removed), Impressions	**0.7695**	0	0
4	Write-up (items treated individually)	0.3952	1	0

[1]moderate mean and low item/total-test-score correlation
[2]moderate mean, low item/total-test-score correlation, and alpha-if-item-deleted increases
Bold alpha coefficients indicate final revised subscales

Case C. The initial analysis included all items from both the SP encounter and the student write-up to establish a base estimate of reliability upon which to improve. The initial reliability estimate was $\alpha = 0.6173$. Case C has an SP encounter section but does not have a write-up section. As with Case A, the data suggested that History and Physical Exam were measuring two different constructs. Unlike Case A, however, the History component did not seem to break down into two subcomponents. An analysis was run on the same items that made up the first History component in Case A, items 1 through 7 excepting 4, but reliability dropped dramatically to $\alpha = 0.3568$, compared to a reliability of $\alpha = 0.7175$ for the same items in Case A.

The items were broken down in the more intuitive sense. Introductory and History items, 1 through 14, were grouped and rerun as one subscale; Physical Exam and Impressions items, 15 through 26, were grouped and rerun as a second subscale. The first

subscale generated a reliability estimate of $\alpha = 0.5213$. Two items were identified as meeting the criteria for removal: item 4 and item 12. These items had moderate *p* values, low item/total-test-score correlations, and an alpha-if-item-deleted that suggested reliability would increase if the item was removed. Items 4 and 12 were removed and the estimated reliability increased to $\alpha = 0.5897$. Item 14 then met the criteria for removal. Item 14 was removed and reliability estimate increased to $\alpha = 0.5985$.

A reliability analysis of the Physical Exam items was performed. Items 19 and 26 were identified as problematic. Item 26 is an Impressions item. This item in Case A and Case B served to increase reliability; however, in Case C it did not seem to fit with the Physical Exam items and was removed. Item 19 was also removed. The resulting reliability estimate for the Physical Exam section was $\alpha = 0.6436$.

Table 7 shows the steps described above for Case C. The proposed structure for Case C includes a History component and a Physical Exam component treated as separate scales. Items 4, 12, and 14 were removed from the first subscale and items 19 and 26 were removed from the second subscale. The original scale had five items with both moderate *p* values and low item/total-test-score correlations. The revised subscales have none.

Case D. Case D was comprised entirely of a write-up. The initial reliability analysis of Case D included all items from the write-up to establish a base estimate of reliability upon which to improve. There was a History component, a Diagnosis component, and an Ethics component. An analysis run on all items generated the lowest reliability of all four cases, $\alpha = 0.3193$.

Table 7: Summary of Iterations for Case C Estimated Reliability Improvement

Iteration	Items Included	Alpha Coeff.	# of Problematic Items[1]	# of Items for Removal[2]
1	All items 1-26 – Introductory, History, Physical Exam, Impressions	0.6173	5	1
2	Introductory, History	0.5213	2	2
3	Introductory, History (items 1-14, items 4 and 12 removed)	0.5897	1	1
4	Introductory, History (items 1-14, items 4, 12, and 14 removed)	**0.5985**	0	0
5	Physical Exam, Impressions	0.6250	2	2
6	Physical Exam (items 15-24, item 19 removed), Impressions (items 25-26, item 26 removed)	**0.6436**	0	0

[1]moderate mean and low item/total-test-score correlation
[2]moderate mean, low item/total-test-score correlation, and alpha-if-item-deleted increases
Bold alpha coefficients indicate final revised subscales

A second analysis was run on History, Diagnosis, and Ethics items in three separate groups. Alpha coefficients dropped on both History and Diagnosis $\alpha = 0.2881$ and $\alpha = 0.2307$ respectively, but increased in Ethics to $\alpha = 0.3826$.

A third analysis combined the History and Diagnosis components and the alpha coefficient decreased to $\alpha = 0.2918$. However, item 6 was identified for removal with a moderate *p* value, a low item/total-test-score correlation, and alpha-if-item-deleted showing that reliability would increase if the item was eliminated. A final analysis was run having eliminated item 6. The reliability estimate increased to $\alpha = 0.3207$.

Table 8 shows the steps described above for Case B. The proposed structure for Case D includes two subscales appearing to measuring two separate constructs. A component containing History and Diagnosis items with item 6 removed, and an Ethics

component. The original scale had two items with both moderate *p* values and low item/total-test-score correlations. The revised subscales have only one between them.

Table 8: Summary of Iterations for Case D Estimated Reliability Improvement

Iteration	Items Included	Alpha Coeff.	# of Problematic Items[1]	# of Items for Removal[2]
1	All items 1-11 – History, Diagnosis, Ethics	0.3193	2	1
2	Ethics	**0.3826**	0	0
3	History, Diagnosis	0.2918	2	1
4	History, Diagnosis (items 1-7, item 6 removed)	**0.3207**	1	0

[1]moderate mean and low item/total-test-score correlation
[2]moderate mean, low item/total-test-score correlation, and alpha-if-item-deleted increases
Bold alpha coefficients indicate final revised subscales

Correlations between item and "overall" rating. Each medical case was assigned an overall rating. The overall rating was a holistic evaluation assigned to each student based on his or her performance on a particular medical case. A correlation was run between the overall rating and each individual item. The strongest correlations appeared between the overall rating and the write-up items and the Impressions items. This was most apparent in case A where statistically significant correlations were identified in nine out of 30 items: four out of five write-up items had significant correlations, as did both Impressions items. These six items had the highest correlations ranging from 0.240 to 0.425. The remaining three items with significant correlations had correlations ranging from 0.108 to 0.134. A similar pattern emerged in Case B. Of the 21 items, 13 had significant correlations with the overall rating, including all four write-up items and both Impressions items. The correlations ranged from 0.173 to 0.625. The other seven items with significant correlations ranged from 0.106 to 0.264.

Case C showed the most comprehensive distribution of significant correlations indicating that the majority of the items were contributors to the overall rating. Of the 26 total items, 19 had statistically significant correlations. Case C did not have a write-up section; still, the highest correlations resided with the Impressions items at 0.374 and 0.485. The significant correlations for the remaining items ranged from 0.109 to 0.255. Case D was comprised solely of a write-up section and all but one of the items correlated significantly with the overall rating. Significant correlations ranged from 0.118 to 0.366. Table 9 summarizes these results. A complete documentation of the correlations between all items and overall rating can be found in Appendices F, G, H, I for Cases A, B, C, and D respectively.

Table 9: Summary of Statistically Significant Correlations between Item Scores and Overall Rating

	CASE A		CASE B		CASE C		CASE D	
Item Type	**No. of Items**	**Corr. Low High**	**No. of Items**	**Corr. Low High**	**No. of Items**	**Corr. Low High**	**No. of Items**	**Corr. Low High**
Write-Up Items	4 of 5	0.253 0.425	4 of 4	0.173 0.625	N/A	N/A	10 of 11	0.118 0.366
Impressions Items	2 of 2	0.240 0.319	2 of 2	0.202 0.243	2 of 2	0.374 0.485	N/A	N/A
Other Items	3 of 23	0.108 0.134	7 of 16	0.106 0.264	17 of 24	0.109 0.255	N/A	N/A
Correlations are significant at the $p < 0.05$ level								

Substantive Analyses

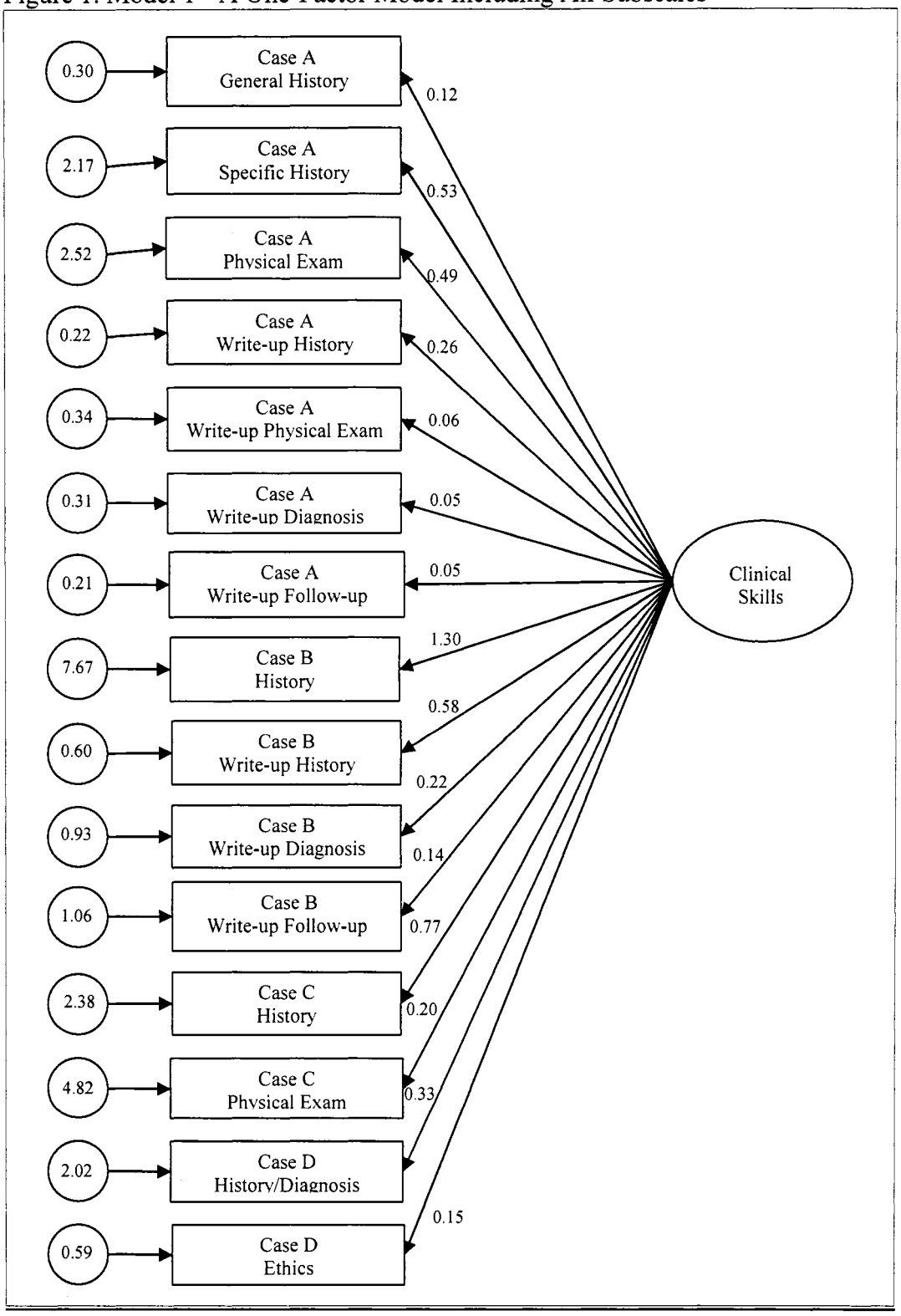
The subscales that emerged from the structural analysis were identified, and a confirmatory factor analysis (CFA) was run to explore whether the subscales were measuring a single dimension or multiple dimensions. The revised subscales included in the one-dimensional model from Case A were General History, Specific History,

Physical Exam, Write-up History, Write-up Physical Exam, Write-up Diagnosis, and Write-up Follow-up Plan. The revised subscales from Case B were History, Write-up History, Write-up Diagnosis, and Write-up Follow-up Plan. Subscales included from Case C were History and Physical Exam. Finally, subscales included from Case D were Write-up History/Diagnosis and Write-up Ethics. This model loaded all the subscales on a single factor, or latent variable, called Clinical Skills. The Legibility subscale from the Write-up section of Cases A and B was eliminated from the analysis as it was not hypothesized to contribute from a model building perspective. This one-dimensional model is represented in Figure 1.

Model parameter statistics were generated for this one-dimensional model. Factor loadings on Write-up subscales ranged from 0.050 to 0.583; factor loadings on the standardized patient (SP) subscales ranged from 0.115 to 1.30. Four loadings were found to be not statistically significant, three of which were Write-up subscales. The standard error for each Write-up variable ranged from 0.207 to 2.016 and for each SP variable ranged from 0.303 to 7.674. All were statistically significant.

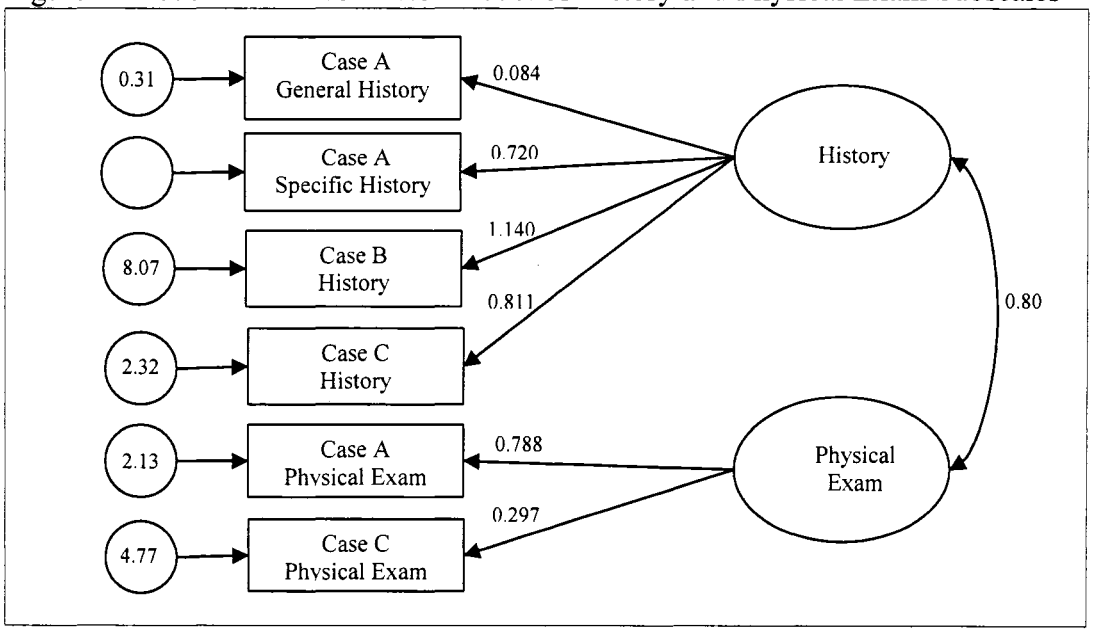
Overall fit statistics were generated. The chi-square (χ^2), the root mean square error of approximation (RMSEA), and the comparative fit index (CFI) values were generated. The overall fit statistics for this one-dimensional model were $\chi^2(90) = 170.399$, $p \leq 0.01$, RMSEA = 0.051, with the 90% confidence interval from 0.039 to 0.063, and CFI = 0.684.

Figure 1. Model 1 - A One-Factor Model Including All Subscales



Another model was proposed which hypothesized that the subscales that emerged from the structural analysis were measuring two primary dimensions rather than one. The subscales represented in this model from Case A were General History, Specific History, and Physical Exam. The Physical Exam subscale from Case B was included, as were the History and Physical Exam subscales from Case C. Subscales comprised of Write-up items were excluded from this model. This two-dimensional model is represented in Figure 2.

Figure 2. Model 2 - A Two-Factor Model of History and Physical Exam Subscales



Model parameter statistics for this two-dimensional model were generated. Factor loadings for the variables ranged from 0.084 to 1.140. Three of the six factor loadings were found to be not statistically significant. The standard error for each variable ranged from 0.309 to 8.072. All were statistically significant. Overall fit statistics for this two-dimensional model were χ^2 (8) = 8.777, p = 0.361, RMSEA = 0.017 with the 90%

confidence interval 0.0 - 0.067, and CFI = 0.989. Table 10 summarizes the statistics for the two models.

Table 10. Summary of Model Statistics

	One-Factor Model	Two-Factor Model
Factor loading range	0.050 to 0.768	0.084 to 0.811
Measurement Error Variance Range	0.207 to 7.674	0.309 to 8.072
Chi-square	170.399, $p \leq 0.01$	8.777, $p = 0.361$
RMSEA	0.051	0.017
CFI	0.684	0.989

A third model looked at all subscales fitted into the four cases. This model was a poor fit to the data, suggesting that a single score for each case may not be an accurate representation of students' performance.

External Analyses

The scores from the revised subscales were correlated with USMLE Step 2 CK scores. The USMLE Step 2 CK and the OSCE scores were not expected to correlate well as they measure different constructs. This was found to be true. Case A had three new subscales and five items to be treated individually. Only the write-up Physical Exam item correlated with the USMLE Step 2 CK scores at a level of statistical significance ($r = 0.163, p < 0.01$). Case B had one scale and four items to be treated individually. Only the write-up Follow-up item correlated with the USMLE Step 2 CK scores at a level of statistical significance ($r = 0.119, p < 0.05$). Case C had two subscales. The Physical Exam subscale correlated with the USMLE Step 2 CK scores at a level of statistical significance ($r = 0.186, p < 0.01$). Case D had two subscales. The History/Diagnosis subscale correlated with the USMLE Step 2 CK scores at a level of statistical significance

($r = 0.115, p < 0.05$). Appendix J summarizes the relationships between USMLE Step 2 CK and each subscale.

This chapter has outlined the results of the analyses. It has described the outcomes of the structural, substantive and external analysis as sources of evidence of OSCE score validity. The next chapter presents a discussion of these results.

Chapter 5: Discussion

This study looked at four Objective Structured Clinical Examination (OSCE) medical cases to explore issues of score reliability and validity. The OSCE is administered to ensure that medical students are prepared to move on to the next level of clinical skills training and to provide formative instruction to the students. This study explored whether the checklist items from the different OSCE cases were truly measuring one dimension that could be summarized by a single overall rating, or if there were potentially multiple dimensions and, therefore, multiple subscales. It also explored whether all items contributed to the overall rating and accurately reflected the clinical skills proficiency of the student. Finally, it explored how well the scores from these refined subscales, if indeed they existed, might predict judgments about future student performance.

This chapter opens with a discussion of the findings with respect to evidence of construct validity. The analyses revealed that the medical case score interpretations may achieve greater reliability and validity if some items were removed and if subscales were created. Second, this chapter describes the limitations of the study. This chapter closes with directions for future research.

Review of the Evidence of Construct Validity

In performance assessment, as with any assessment, it is important that the score reflect the level of ability or skill being measured. Studies of score reliability and validity explore whether the score interpretation is reasonable representation of the level of performance. In the OSCE at the University of Washington, the students perform many

tasks – they illicit a patient history, maybe perform a physical exam, make a diagnosis, and perhaps write up their findings – but only receive a single overall rating of exceeds expectations, meets expectations, or needs development. It is possible that each checklist contains a substantial amount of information that is ultimately buried in a single score. For instance, one student may do poorly on the History component across cases but do well on the Physical Exam component, while another student may perform poorly on Physical Exam component but well on the History component. These dissimilar performances may, however, receive the same overall rating of "meets expectations." Similarly, one student may do poorly on the History component on all medical cases, but still receive a score of meets expectations on each. The student's difficulty with the History items is not being captured in the overall rating.

The first research question asked whether the checklist items from the different OSCE cases were truly measuring one dimension that could be summarized by a single overall rating, or if there were potentially multiple dimensions and, therefore, multiple subscales. It also asked about the internal consistency of these refined scales and subscales. An assessment – in this case, a checklist – that can identify that different dimensions are being measured can provide valuable formative feedback to the student in the form of multiple scores instead of a single overall rating. A checklist that can discriminate between levels of performance can provide valuable feedback to the student, especially to the students that are in the "borderline" range of performance. Studies of the reliability and validity of assessment scores attempt to provide evidence that supports the interpretation of such information. When an assessment shows signs of internal

consistency and signs of discriminating between levels of performance there is evidence that the score interpretation is valid for its intended purpose.

An assessment that is covering a single dimension will show signs of higher internal consistency among the items. Coefficient alpha provides internal consistency reliability information. Item difficulty and item discrimination statistics can be used to identify items that are detracting from the internal consistency, and also to regroup items so that one dimension is being measured in each scale. In both these cases, internal consistency would increase. It is unwise from a psychometric standpoint to assign an overall rating to a group of items that show evidence of measuring different dimensions of a performance. It was with this in mind that the analysis of the scores assigned to the OSCE cases was approached.

It is important to acknowledge that items were deleted from the reliability analyses to demonstrate the deleterious effect of problematic items on the reliability and validity of the scale and to demonstrate how more sound measurement scales and subscales could be achieved. In practice, assuming they measured valid aspects of clinical skills and should be included, these items would be brought to the attention of the test designer so they could be modified and improved. Modifying and improving the items may result in one of several different outcomes. The modified item may fit better with the existing scales. The revised items may cause some movement between subscales, or entirely new subscales may be created. Lastly, each improved item may be treated individually. The checklist developers could then use the information obtained from unique items, without adding noise to a scale. For example, an item that assesses whether

the examinee asked the patient for a history of the present illness may be an item that every examinee gets correct. From a psychometric standpoint, this item adds no measurement value; however, that every student got this item correct may be valuable information for the examiners. Any of these outcomes would likely have a positive impact on score reliability. The remainder of this section of the chapter discusses how the items in each of the OSCE cases were explored and how more reliable scales and subscales were achieved.

Content Evidence of Construct Validity

The cases were designed by physician-educators considered experts in their respective fields. The cases were reviewed by their peers for evidence of both construct representation and construct-irrelevant variance. This evidence of face validity was strongly considered as the structural analysis unfolded. As items were considered for revision or deletion it was noted that the items may need to be included for adequate domain coverage and for didactic purposes regardless of whether the statistical analyses supported the items' inclusion.

Structural Evidence of Construct Validity

Item difficulty. While it was expected that item means would be higher than the norm-referenced range of 0.30 to 0.70 with a 0.50 average, it was not expected that the means for the same item would vary so dramatically across cases. One possible explanation is that the same items are not appropriate for all the medical cases. For example, item 13 had p values that ranged from 0.0665 to 0.6821. It is a patient history question regarding the patient's dietary habits. It could be that dietary habits are much

less relevant to Case A; therefore, if the examinee already has a sense of the diagnosis, he or she does not ask about dietary habits specifically and, therefore, does not receive credit for the item. Dietary habits may be much more relevant to Case B and, therefore, the examinees are much more likely to specifically seek that aspect of patient history. Research shows that physicians with more experience may be able to skip items that are not necessary to make the diagnosis (Cerilli, Merrick, & Staren, 2001). It may be that this phenomenon is emerging at this stage of physician development. If the students are taught to ask the same common items regardless of the medical case, clearly the students are not getting this message during their training. If, however, the students are taught to ask of patients only questions relevant to that particular medical case, then the checklist should be redesigned to reflect only relevant items.

Item discrimination. These values were generally relatively low. Item discrimination values provide the most information about the ability of the student when they fall within the 0.30 to 0.70 range. These relatively low item discrimination scores indicated that the students that were answering an item correctly were not necessarily the same students that received a high score on the case overall. The item discrimination values across all four medical cases ranged from a low of r_{iX} = -0.1089 to a high of r_{iX} = 0.5310. The revised scales and subscales were comprised of items with overall higher item discrimination values. These values are still not as high as is recommended for gaining the most information about the different levels of performance; however, these increased values demonstrate that the items are better able to discriminate between levels

of performance. This, in turn, increases the reliability and validity of the score interpretations.

Case A. Case A consisted of 30 items including Introductory, History, Physical Exam, and Impressions items. The examinees received an overall rating of exceeds expectations, meets expectations, or needs development.

An analysis of the instrument at the item level revealed that there may be four individual subscales. The first subscale composed of six checklist items includes the two Introductory items and four History items. A fifth History item was removed as it did not fit with any of the subscales. These six items could be categorized as the most general introductory type items that may be asked regardless of patient ailment. This subscale was labeled "General History." The second subscale was composed of the remaining eight History items. These items may also be asked regardless of patient ailment; however, they begin to ask for more specific information. This subscale was labeled "Specific History." The third subscale was comprised primarily of six Physical Exam items. An additional two Physical Exam items were removed from the subscale as they did not contribute to the reliability of the scale. The final two items in this subscale were the Impressions items of organization/flow and communication skills. These two items when grouped with other subscales reduced the reliability of each subscale. Their fit with the Physical Exam items suggests that organization and communication are important to the physical exam process for Case A.

The final subscale that consisted of items from the write-up section did not appear to be measuring a single dimension of performance. Two of the five write-up items

reflect what the examinees learned in the History and Physical Exam components of the SP encounter. The History and Physical Exam items on the SP encounter checklist suggested that two different dimensions were being measured; therefore, it is not surprising that the write-up items behave similarly. Additionally, a third of the five write-up items assessed legibility, a quality unlikely to group with any other item. These items are important from the standpoint of providing content related evidence of score validity as test designers are generally experts in their field and able to judge appropriate item inclusion. However, because these items do not create a sound subscale from a psychometric standpoint, they are items that may best be looked at individually rather than as a group of items measuring the same dimension.

Overall, Case A consisted of multiple dimensions: General History, Specific History, Physical Exam, and the write-up items. These subscales appeared to better reflect the ability of the examinees and to better reflect the dimensions being covered, both of which suggested increased score validity.

Case B. Case B originally consisted of 21 items for which the student received an overall rating of exceeds expectations, meets expectations, or needs development. Case B involved both an SP encounter section and a write-up section, but, unlike Cases A and C, the SP encounter section did not include a Physical Exam component. An analysis of the instrument at the item level revealed a single dimension; however, the analysis revealed several items to be removed or to be treated individually and not as part of any scale. The scale included both Introductory items, all but two History items, and both Impressions items. Unlike Case A, the History items did not group effectively into two different

subscales. When the History items were separated in a manner comparable to Case A, the reliability estimates of each subscale were dramatically lower. The two items removed from the analysis were items that were also identified as problematic in Case C. As with Case A, the four write-up items seemed to be best treated individually.

Case C. Case C originally consisted of 21 items for which the student received an overall rating of exceeds expectations, meets expectations, or needs development. As with Case A and Case B, the data suggested that a more valid score interpretation may be obtained by breaking the single scale into subscales. As with Case B, the History items did not appear to belong on two separate subscales. Rather, the History items seemed to form a single scale. As with Case A, item 4 was identified as problematic and was removed from the analysis. Item 12 was also removed. The second subscale was comprised of Physical Exam items and Impressions items. In this subscale, item 19 was removed from the analysis, as was the communications skills item. In practice, these items may be important for evidence of content validity and would remain in the scale. It may be that the items need to be reworded, the physician-evaluator needs to be further trained, or the examinees need to be instructed differently to ensure that they understand how communication differs from case to case. Whatever the solution, the data suggest that these items require attention.

Case D. Case D consisted solely of a write-up section comprised of 11 items for which the examinee received an overall rating of exceeds expectations, meets expectations, or needs development. It was originally hypothesized that three subscales, a History scale, a Diagnosis scale, and an Ethics scale might emerge as suggesting greatest

score validity. It was found, however, that the three History items and the three Diagnosis items created a single subscale, which makes sense given that diagnosis is tightly tied to history taking. A Diagnosis item was removed for greatest estimates of reliability and the remaining six items formed the "History/Diagnosis" subscale. The Ethics items grouped onto a second subscale. The reliability for each of these subscales was greater than the original unitary scale.

Across Cases. A comparison of items across cases identified several items as problematic on multiple medical cases. First, items 3 and 4, were the two items categorized as History of Present Illness. They were observed over Cases A, B, and C. They appeared to be linked to one another and had problematic data. Item 3 asked whether the examinee obtained a history of the present illness. In all the medical cases, the vast majority of examinees did this. In Case B, item 3 was removed because all the students performed the task and the item, therefore, provided no measurement information. In practice, the item may remain in the scale from a content validity perspective; however, the item does not add measurement value to the scale. Item 4 asked whether the examinee asked the patient's thoughts regarding the cause of the symptoms. In Cases A, B, and C, the proportions of students asking this of the standardized patient were moderate. The item/total-test-score correlations were very low for this item indicating that the examinees getting credit for the item were not necessarily the examinees that did well on the case overall. Furthermore, the estimated reliability of the scale overall would have increased if item 4 were removed. Therefore, item 4 was removed from the scales in Cases A and C to create a more reliable scale. One possible

explanation is that the examinee feels that he or she obtains the answer to item 4 from the prior item; therefore, the examinee does not ask item 4 outright and, as a result, does not receive credit for it. Whatever the reason, it is important that examiners reevaluate these two items.

Two other items also appear to be linked and were identified as problematic in more than one case. One of the items is checked if the student asks about a history of this illness in the family. The other is checked if the examinee asks about marriage, children, etc. It is possible that these items were overlapping in the viewpoint of the examinees and/or the physician-evaluator. In both Case B and Case C the social history item was removed as it was identified as having a moderate mean, a low item/total-test-score correlation, and the case reliability would increase if the item were removed. Although the family history item was not removed from Case A, it did show signs of being problematic as it had a moderate mean and a low item/total-test-score correlation. Again, in reality, these items would probably not be eliminated as they may be important for content validity evidence. What is important is that these items are brought to the attention of the examiners for further exploration.

Substantive Evidence of Construct Validity

A confirmatory factor analysis was performed to test the theory that the OSCE checklists are a two-dimensional assessment and, therefore, assigning a single score may not result in a valid score interpretation. Assigning a single score to a performance on an OSCE case suggests that a single dimension is being measured. This single score is expected to capture the level of performance and feedback to the student comes in this

single score form. Based on this presumption, the first model included all the subscales that emerged from the analysis of the internal structure of the OSCE checklist, except for the Legibility subscale as it was not a variable of interest. The factor loadings for the Write-up variables were low and of the four non-significant variable loadings, three belonged to Write-up items. This supported the structural analysis which suggested that while the Write-up items did not group with the History or Physical Exam subscales they also did not form their own subscale and should be treated separately. The measurement error of the variables was quite high for many of the variables which may be explained by both misfit to the model and the low reliability inherent in the subscales themselves.

The chi-square of 170.399 with $p \leq 0.01$ suggested rejection of the hypothesis that the proposed one-factor model fits the covariance patterns of the observed data. The root mean square error of approximation (RMSEA) which estimates that lack of fit between the hypothesized model and the observed data was acceptable at RMSEA = 0.051. The comparative fit index (CFI) was low at CFI = 0.684. As hypothesized, there were several statistical indicators suggesting that the model representing a single dimension may not be the best fit for the observed OSCE scores.

Based on theory and the results of the empirical analysis of the internal structure of the measurement instrument, the second model reflected the possible existence of two factors: a History factor and a Physical Exam factor on to which the respective history and physical exam subscales from the medical cases would load. This model removed the Write-up components, as all the evidence suggested that these were the least stable and should be treated as separate subscale scores. The chi-square value for the two-dimension

model dropped to 8.777 with $p = 0.361$. This indicated non-rejection of the hypothesis that the two-factor model fit the original data covariance patterns. As compared to the one-factor model, the RMSEA dropped to RMSEA = 0.017 and the CFI improved to CFI = 0.989. As hypothesized, there were several statistical indicators that the two-factor model better fit the data than the one-factor model.

Stevens (2002) states the following about decision making on statistical grounds alone:

> It is essential that the researcher be able to base a model on theory...because...it is not always possible to distinguish between different models on statistical grounds alone. In many cases, theoretical considerations are the only way in which one model can be distinguished from another. (p. 415)

The improvement of the two-factor model over the one-factor model was dramatic enough to reasonably suggest that the statistical data supports the hypothesis of a better fitting two-factor model. However, for equitable comparison of models, a third factor analysis was run on a one-factor model that loaded only the subscales included in the two-factor model. The statistical differences between these two models were negligible. Therefore, theoretical considerations must be relied upon to distinguish between the two two-factor models.

In summary, the results suggest that the OSCE does measure multiple dimensions. They also suggest that internal consistency is improved when these dimensions are considered.

Item Correlations with Overall Rating

The second research question asked whether the overall rating was an accurate representation of the clinical skills proficiency of the student and whether all items contributed equally. With the emergence of multiple subscales measuring multiple dimensions, the data suggested that they were not contributing equally to overall rating. The correlations between each item and the overall rating assigned to the medical case were explored. In the medical cases where there existed a write-up section, the overall rating was assigned by the physician-evaluator who scored the write-up. The physician-evaluator that scored the write-up and assigned the overall rating was not the same physician-evaluator that observed and scored the SP encounter. In assigning the score, the physician-evaluator was to take into account both the standardized patient encounter checklist and his or her own write-up checklist scores. The results of these analyses suggest that the strongest correlations occurred between the write-up items and the overall rating, and the Impressions items and the overall rating. This suggests that the standardized patient encounter is not being weighted as heavily in the overall rating as the write-up or the Impressions items. To discount or underestimate the scores from the standardized patient encounter limits the range of the domain and may be a source of construct under-representation. In this case, the score assigned may not represent the ability of the student and the validity of the score inference may be in compromised.

External Evidence of Construct Validity

The third research question asked how scores from the revised subscales may predict performance on other medical school assessments. To seek evidence of construct

validity through external sources, correlations between the new subscales and USMLE Step 2 Clinical Knowledge (CK) were generated. The OSCE and the USMLE Step 2 CK measure different domains and, therefore, low correlations between the two sets of scores were expected. Although several of the subscales did correlate with USMLE Step 2 CK scores at a statistically significant level, these correlations were not high enough to be valuable in a practical sense. These low correlations provided discriminant evidence, as we expect measures of differing domains to share low correlations.

Review of Findings

The OSCE serves many valuable purposes in clinical skills assessment. One of those is identifying students for remediation who are not fully prepared to transition to residency. It is likely that an OSCE can accomplish this with only limited success without high reliability coefficients and without significant evidence of score validity. However, students report great satisfaction with the OSCE experience (Scott, 2006). It provides a structure for their study time, and exposes them to what they can expect to experience as a future clinical physician. Many students would like to opportunity to participate in more OSCE administrations. The OSCE is clearly a valuable teaching tool for these reasons alone. It is possible, however, that the OSCE can provide increased value through improved discrimination between performance levels as well as improved feedback to the students for formative development.

There is evidence that analyzing the instrument at the item level could be a way to more clearly discriminate between ability levels and become an even more valuable teaching tool and assessment instrument. The item analysis also serves to identify items

for modification or removal which can positively contribute to the next iteration of OSCE checklists. As checklists are further refined, instructors will be better able to identify the various levels of ability and better identify those students that needed development in some areas rather than just being able to identify the most dramatic outliers. It is important that the students not see the OSCE as something that "everyone passes," as that view may negatively influence the amount of time committed to preparing for them. If the OSCE scores are not demonstrated to be reliable and valid for the purpose of identifying clinical skills competence, educators will be reluctant to judge an examinee's performance based on them, and even less willing to take action.

It is especially important that the items are able to distinguish between levels of performance around the cut off score. Although many students may clearly exceed expectations on each case and others may clearly need development, there are other students that may not fall so clearly into one classification or the other. They may perform at a borderline level across medical cases and medical case components; alternatively, they may excel in one area but perform poorly in another. The validity of the score interpretations is especially important for these students. To inaccurately assign an examinee a score of exceeds expectations or meets expectations when the student actually needs development may allow a student to move onto the next step of the program when she or he is unprepared. Such a student may be deprived of valuable remediation.

Performance assessment is an inexact science. There are rarely right or wrong answers; rather there are arguments for and against, and evidence to support or deny,

score interpretations and methods. This study contributes to the already existing body of literature regarding the reliability and validity issues surrounding the OSCE. The intent was to further understanding of reliability and validity issues by examining the technical quality of the OSCE at the item level. It summarized a review of the literature that included the approach to validity that is commonly held today, a discussion of reliability and validity studies that surround the OSCE, and the theory that was used to explore the OSCE at the item level. This study also demonstrated a methodology that could be employed to improve existing checklists for OSCE.

This study was not intended to provide hard and fast answers about the ideal OSCE checklist. The study provided a demonstration of ways to improve the OSCE and to consider dimensions of performance formatively in providing feedback to students. Ultimately, the goal is to ensure, to the extent possible, that the score that is assigned to the student's performance is an accurate representation of the clinical skills proficiency of the student. The goal of this study was to explore ways to increase confidence that this is indeed the case.

Limitations

Some limitations to this study must be considered. This study involved a single institution and only two cohorts of students. The small number of medical cases involved in the study may have impacted the results of the analyses. Several issues of consistency surrounded the raters. The level of rater training and experience in OSCE evaluation, and the level of rater knowledge and expertise in medicine varied across raters. A different

rater scored the standardized patient encounter than assigned the overall rating. For this study, there was no way to control for or to assess the rater training and rater effects.

Future Research

Future research may build upon this study in various ways. Including additional cohorts of student data as it becomes available will increase the sample size and allow for more reliable estimates. A larger sample of examinees will also allow for the application of other theories, such as Item Response Theory, for the analyses of the OSCE checklists at the item level. Future studies may also incorporate the effects of increased rater training on the reliability and validity of the OSCE scores.

Future studies may explore other sources of validity evidence. Consequences of the OSCE scores, such as the impact of remediation, may also be explored. Construct validity evidence may also be sought through the ability of the OSCE scores to generalize across various student populations, across programs, across professional schools, and across medical cases. Convergent external evidence of construct validity may be sought through analyses involving residency performance evaluations or an OSCE administered during residency. USMLE Step 2 Clinical Skills scores would also be a source of convergent evidence for medical schools whose student scores contain adequate variability. These possibilities for future research studies will further understanding of the reliability and validity issues surrounding the OSCE.

List of References

AERA APA and NCME. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Allen, M. J., & Yen, W. M. (2002). *Introduction to Measurement Theory.* Prospect Heights: Waveland Press, Inc.

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 19-32). Hillsdale: Lawrence Erlbaum Associates, Inc.

Auewarakul, C., Downing, S. M., Jaturatamrong, U., & Praditsuwan, R. (2005). Sources of validity evidence for an internal medicine student evaluation system: an evaluative study of assessment methods. *Medical Education, 39*(3), 276-283.

Auewarakul, C., Downing, S. M., Praditsuwan, R., & Jaturatamrong, U. (2005). Item analysis to improve reliability for an internal medicine undergraduate OSCE. *Adv Health Sci Educ Theory Pract, 10*(2), 105-113.

Berk, R. A. (1986). Preface. In R. A. Berk (Ed.), *Performance assessment: methods and applications* (pp. ix-xiv). Baltimore: The Johns Hopkins University Press.

Carraccio, C., & Englander, R. (2000). The Objective Structured Clinical Examination: A Step in the Direction of Competency-Based Evaluation. *Arch Pediatr Adolesc Med, 154*(7), 736-741.

Cerilli, G. J., Merrick, H. W., & Staren, E. D. (2001). Objective Structured Clinical Examination technical skill stations correlate more closely with postgraduate year level than do clinical skill stations. *Am Surg, 67*(4), 323-326; discussion 326-327.

Cohen, D. S., Colliver, J. A., Marcy, M. S., Fried, E. D., & Swartz, M. H. (1996). Psychometric properties of a standardized-patient checklist and rating-scale form used to assess interpersonal and communication skills. *Acad Med, 71*(1 Suppl), S87-89.

DuBois, P. H. (1970). *A History of psychological testing.* Boston: Allyn and Bacon, Inc.

Dugan, R. E. (2006). Stakeholders of higher education institutional accountability. In P. Hernon, R. E. Dugan & C. Schwartz (Eds.), *Revisiting outcomes assessment in higher education* (pp. 39-62). Westport: Libraries Unlimited.

Education Commission for Foreign Medical Graduates. (2007). ECFMG 2007 Information Booklet. Retrieved February 24, 2007, from http://www.ecfmg.org/2007ib/ibcert.html#validity

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 105-146). New York: National Council on Measurement in Education and American Council on Education.

Flexner, A. (1910). *Medical education in the United States and Canada: A report to the Carnegie Foundation for the Advancement of Teaching*. New York City.

Gerrow, J., Murphy, H., Boyd, M., & Scott, D. (2003). Concurrent validity of written and OSCE components of the Canadian dental certification examinations. *J Dent Educ., 67*(8), 896-901.

Giordano, G. (2005). *How testing came to dominate American schools: The history of educational assessment*. New York: Peter Lang Publishing, Inc.

Haladyna, T. M. (1994). *Developing and validating multiple-choice test items*. Hillsdale: Lawrence Erlbaum Associates, Inc.

Harden, R. M., Stevenson, M., Downie, W. W., & Wilson, G. M. (1975). Assessment of clinical competence using objective structured examination. *Br Med J, 1*(5955), 447-451.

Hodges, B., & McIlroy, J. H. (2003). Analytic global OSCE ratings are sensitive to level of training. *Med Educ, 37*(11), 1012-1016.

Hodges, B., Regehr, G., McNaughton, N., Tiberius, R., & Hanson, M. (1999). OSCE checklists do not capture increasing levels of expertise. *Acad Med, 74*(10), 1129-1134.

Humphris, G. M., & Kaney, S. (2001). Examiner fatigue in communication skills objective structured clinical examinations. *Medical Education, 35*(5), 444-449.

Jones, G., Jones, B. D., & Hargrove, T. Y. (2003). *The unintended consequences of high-stakes testing*. Lanham: Rowman & Littlefield Publishers, Inc.

Joorabchi, B. (1991). Objective structured clinical examination in a pediatric residency program. *Am J Dis Child, 145*(7), 757-762.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: The Guilford Press.

Martin, I. G., & Jolly, B. (2002). Predictive validity and estimated cut score of an objective structured clinical examination (OSCE) used as an assessment of clinical skills at the end of the first clinical year. *Med Educ, 36*(5), 418-425.

Martin, J. A., Reznick, R. K., Rothman, A., Tamblyn, R. M., & Regehr, G. (1996). Who should rate candidates in an objective structured clinical examination? *Acad Med, 71*(2), 170-175.

Mazor, K. M., Ockene, J. K., Rogers, H. J., Carlin, M. M., & Quirk, M. E. (2005). The relationship between checklist scores on a communication OSCE and analogue patients' perceptions of communication. *Adv Health Sci Educ Theory Pract, 10*(1), 37-51.

McGraw, R. C., & O'Connor, H. M. (1999). Standardized patients in the early acquisition of clinical skills. *Med Educ, 33*(8), 572-578.

McKinley, D. W., & Boulet, J. R. (2005). Using factor analysis to evaluate checklist items. *Acad Med, 80*(10 Suppl), S102-105.

McLaughlin, K., Gregor, L., Jones, A., & Coderre, S. (2006). Can standardized patients replace physicians as OSCE examiners? *BMC Med Educ, 6*, 12.

McLean, J. E., & Lockwood, R. E. (1996). *Why we assess students - and how: the competing measures of student performance*. Thousand Oaks: Corwin Press, Inc.

Merrick, H. W., Nowacek, G., Boyer, J., & Robertson, J. (2000). Comparison of the objective structured clinical examination with the performance of third-year medical students in surgery. *The American Journal of Surgery, 179*(4), 286-288.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: National Council on Measurement in Education and American Council on Education.

Messick, S. (1992). *The interplay of evidence and consequences in the validation of performance assessments*. Princeton: Educational Testing Service.

Messick, S. (1994). *Alternative modes of assessment, uniform standards of validity*. Princeton: Educational Testing Service.

Messick, S. (1999). The changing face of higher education assessment. In S. Messick (Ed.), *Assessment in higher education* (pp. 3-7). Mahwah: Lawrence Erlbaum Associates, Inc.

National Research Council. (2002). *Performance assessments for adult education, exploring the measurement issues, Report of a workshop*. Washington, D.C.: National Academy Press.

Newble, D. (2004). Techniques for measuring clinical competence: objective structured clinical examinations. *Med Educ, 38*(2), 199-203.

Park, R., Chibnall, J., Blaskiewicz, R., Furman, G., Powell, J., & Mohr, C. (2004). Construct validity of an objective structured clinical examination (OSCE) in psychiatry: associations with the clinical skills examination and other indicators. *Academic Psychiatry [NLM - MEDLINE], 28*(2), 122.

Probert, C. S., Cahill, D. J., McCann, G. L., & Ben-Shlomo, Y. (2003). Traditional finals and OSCEs in predicting consultant and self-reported clinical skills of PRHOs: a pilot study. *Medical Education, 37*(7), 597-602.

Raykov, T. (2004). Estimation of maximal reliability: A note on a covariance structure modelling approach. *British Journal of Mathematical & Statistical Psychology, 57*, 21-27.

Regehr, G., MacRae, H., Reznick, R. K., & Szalay, D. (1998). Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Acad Med, 73*(9), 993-997.

Rifkin, W. D., & Rifkin, A. (2005). Correlation between housestaff performance on the United States Medical Licensing Examination and standardized patient encounters. *Mt Sinai J Med, 72*(1), 47-49.

Rothman, A. I., & Cusimano, M. (2000). A comparison of physician examiners', standardized patients', and communication experts' ratings of international medical graduates' English proficiency. *Acad Med, 75*(12), 1206-1211.

Rothman, A. I., & Cusimano, M. (2001). Assessment of English proficiency in international medical graduates by physician examiners and standardized patients. *Med Educ, 35*(8), 762-766.

Scott, C. (2006). Professor. Seattle.

Sloan, D. A., Donnelly, M. B., Schwartz, R. W., Felts, J. L., Blue, A. V., & Strodel, W. E. (1996). The Use of the Objective Structured Clinical Examination (OSCE) for Evaluation and Instruction in Graduate Medical Education. *Journal of Surgical Research, 63*(1), 225-230.

Stevens, J. P. (2002). *Applied Multivariate Statistics for the Social Sciences* (4th ed.). Mahwah: Lawrence Erlbaum Associates.

Suskie, L. (2006). Accountability and quality improvement. In P. Hernon, R. E. Dugan & C. Schwartz (Eds.), *Revisiting outcomes assessment in higher education* (pp. 13-38). Westport: Libraries Unlimited.

Svinicki, M. D. (2005). Authentic assessment: Testing in reality. In M. V. Achacoso & M. D. Svinicki (Eds.), *Alternative strategies for evaluating student learning*

(Winter 2004 ed., Vol. 100, pp. 23-29). San Francisco: Wiley Subscription Services, Inc.

Taylor, M. L., Blue, A. V., Mainous, A. G., Geesey, M. E., & Basco Jr., W. T. (2005). The relationship between the national board of medical examiners' prototype of the step 2 clinical skills exam and interns' performance. *Academic Medicine, 80*(5), 496-501.

Thistlethwaite, J. E. (2002). Developing an OSCE station to assess the ability of medical students to share information and decisions with patients: issues relating to interrater reliability and the use of simulated patients. *Educ Health (Abingdon), 15*(2), 170-179.

USMLE. (2004). Step 2 clinical skills examination frequently asked questions. Retrieved September 20, 2005, from http://www.usmle.org/FAQs/step2csfaqs1103.htm#structure

USMLE. (2007). 2007 USMLE Step 2 CS Content Description and General Information Booklet. Retrieved April 26, 2007, from http://www.usmle.org/step2/Step2CS/Step2CS2007GI/description.asp

Valentino, J., Donnelly, M. B., Sloan, D. A., Schwartz, R. W., & Haydon, R. C., 3rd. (1998). The reliability of six faculty members in identifying important OSCE items. *Acad Med, 73*(2), 204-205.

van der Vleuten, C. P., van Luyk, S. J., van Ballegooijen, A. M., & Swanson, D. B. (1989). Training and experience of examiners. *Med Educ, 23*(3), 290-296.

Walters, K., Osborn, D., & Raven, P. (2005). The development, validity and reliability of a multimodality objective structured clinical examination in psychiatry. *Medical Education, 39*(3), 292-298.

Wass, V., McGibbon, D., & Van der Vleuten, C. (2001). Composite undergraduate clinical examinations: how should the components be combined to maximize reliability? *Medical Education, 35*(4), 326-330.

Wessel, J., Williams, R., Finch, E., & Gemus, M. (2003). Reliability and validity of an objective structured clinical examination for physical therapy students. *J Allied Health, 32*(4), 266-269.

Wilkinson, T. J., & Frampton, C. M. (2004). Comprehensive undergraduate medical assessments improve prediction of clinical performance. *Medical Education, 38*(10), 1111-1116.

Wilkinson, T. J., Frampton, C. M., Thompson-Fawcett, M., & Egan, T. (2003). Objectivity in objective structured clinical examinations: checklists are no substitute for examiner commitment. *Acad Med, 78*(2), 219-223.

Appendix A: Case A Item Data

CASE A			
Item	**Mean**	**Item/Total-Test-Score Correlation**	**Alpha-if-Item-Deleted**
Item 1 Introductory	0.9697	0.1016	0.5300
Item 2 Introductory	0.9848	0.1676	0.5283
Item 3 History	0.9818	0.2501	0.5244
Item 4 History	0.6333	0.0926	0.5322
Item 5 History	0.9848	0.2066	0.5268
Item 6 History	0.9879	0.1649	0.5289
Item 7 History	0.9727	0.2565	0.5222
Item 8 History	0.6121	0.2202	0.5123
Item 9 History	0.8636	0.1715	0.5215
Item 10 History	0.8061	0.3037	0.5036
Item 11 History	0.9394	0.2512	0.5179
Item 12 History	0.6636	0.1991	0.5159
Item 13 History	0.0667	0.0235	0.5353
Item 14 History	0.7455	0.0645	0.5351
Item 15 History	0.6061	0.2188	0.5126
Item 16 Physical Exam	0.2515	0.2565	0.5083
Item 17 Physical Exam	0.6758	0.1567	0.5224
Item 18 Physical Exam	0.6242	0.2382	0.5095
Item 19 Physical Exam	0.6697	0.1021	0.5306
Item 20 Physical Exam	0.6758	0.2275	0.5116
Item 21 Physical Exam	0.8909	0.0345	0.5356
Item 22 Physical Exam	0.5909	0.0220	0.5433
Item 23 Physical Exam	0.8818	0.1261	0.5266
Item 24 Impressions	1.0394	0.4419	0.4782
Item 25 Impressions	1.2152	0.2116	0.5137
Write-up Legibility	1.5242	-0.1089	0.5670
Write-up History	1.3394	0.2454	0.5073
Write-up Physical Exam	1.0273	0.0891	0.5359
Write-up Diagnosis	1.1364	0.0019	0.5501
Write-up Follow-up Plan	1.1515	0.0083	0.5436
Average of the means n = 330	0.6895	Alpha =	0.5333

Items 1 through 23 are on a 0-1 point scale; all other items are on a 0-2 point scale.

Appendix B: Case B Item Data

CASE B			
Item	**Mean**	**Item/Total-Test-Score Correlation**	**Alpha-if-Item-Deleted**
Item 1 Introductory	0.9855	0.0832	0.6930
Item 2 Introductory	0.8728	0.1044	0.6925
Item 3 History	1.0000		
Item 4 History	0.5260	0.2424	0.6832
Item 5 History	0.9682	0.2775	0.6871
Item 6 History	0.8179	0.3250	0.6786
Item 7 History	0.8960	0.3971	0.6771
Item 8 History	0.6012	0.4540	0.6651
Item 9 History	0.5462	0.4350	0.6665
Item 10 History	0.6618	0.5310	0.6593
Item 11 History	0.4884	0.4819	0.6622
Item 12 History	0.5000	0.4720	0.6631
Item 13 History	0.6821	0.2015	0.6866
Item 14 History	0.9191	0.0947	0.6926
Item 15 History	0.6329	0.1313	0.6923
Item 16 Review of Systems	0.8208	0.2488	0.6836
Item 17 Impressions	1.0173	0.4726	0.6674
Item 18 Impressions	1.0954	0.2493	0.6838
Write-up Legibility	2.9798	0.1532	0.6982
Write-up History	2.6821	0.3337	0.6772
Write-up Diagnosis	2.4624	0.2843	0.6863
Write-up Follow-up Plan	2.2110	0.1142	0.7178
Grand means n = 346	0.6769	Alpha =	0.6924
Items 1 through16 are on a 0-1 point scale; items 17 and 18 are on a 0-2 point scale; all other items are on a 0-4 point scale.			

Appendix C: Case C Item Data

CASE C			
Item	**Mean**	**Item/Total-Test-Score Correlation**	**Alpha-if-Item-Deleted**
Item 1 Introductory	0.9740	0.1031	0.6155
Item 2 Introductory	0.9566	0.0594	0.6175
Item 3 History	0.9769	0.0758	0.6166
Item 4 History	0.4913	0.0801	0.6221
Item 5 History	0.9249	0.1373	0.6129
Item 6 History	0.8410	0.1716	0.6097
Item 7 History	0.8757	0.2483	0.6034
Item 8 History	0.5318	0.1858	0.6087
Item 9 History	0.4884	0.2041	0.6063
Item 10 History	0.7486	0.2719	0.5985
Item 11 History	0.6416	0.3104	0.5928
Item 12 History	0.5694	0.0418	0.6267
Item 13 History	0.9653	-0.0067	0.6204
Item 14 History	0.6763	0.1366	0.6143
Item 15 Physical Exam	0.9855	0.1576	0.6143
Item 16 Physical Exam	0.2948	0.0023	0.6294
Item 17 Physical Exam	0.3064	0.2549	0.6001
Item 18 Physical Exam	0.6474	0.3018	0.5939
Item 19 Physical Exam	0.7023	0.1466	0.6130
Item 20 Physical Exam	0.8439	0.1924	0.6078
Item 21 Physical Exam	0.5347	0.3134	0.5919
Item 22 Physical Exam	0.4133	0.2718	0.5975
Item 23 Physical Exam	0.3815	0.1968	0.6072
Item 24 Physical Exam	0.6705	0.1682	0.6106
Item 25 Impressions	1.0780	0.5058	0.5618
Item 26 Impressions	1.2428	0.2457	0.6009
Grand means n = 346	0.6701	Alpha =	0.6173
Items 1 through 24 are on a 0-1 point scale; items 25 and 26 are on a 0-2 point scale.			

Appendix D: Case D Item Data

CASE D			
Item	**Mean**	**Item/Total-Test-Score Correlation**	**Alpha-if-Item-Deleted**
Item 1 History	0.9451	0.1267	0.3031
Item 2 History	0.7428	0.1431	0.2868
Item 3 History	0.7977	0.0494	0.3206
Item 4 History	0.4075	0.1170	0.2962
Item 5 Diagnosis	0.9509	0.0340	0.3688
Item 6 Diagnosis	0.8295	0.1808	0.2597
Item 7 Diagnosis	1.3179	0.2335	0.2275
Item 8 Ethics	0.9653	0.0875	0.3122
Item 9 Ethics	0.9422	0.0655	0.3146
Item 10 Ethics	0.8584	0.1279	0.2960
Item 11 Ethics	0.3382	0.1365	0.2882
Grand means n = 346	0.6497	Alpha =	0.3193
Items 1 through 4 and 10 through 13 are on a 0-1 point scale; items 6 through 8 are on a 0-2 point scale.			

Appendix E: Case A Correlations between Items and Overall Rating

CASE A	
Item	**Correlation**
Item 1 Introductory	0.108 *
Item 2 Introductory	0.077
Item 3 History	0.073
Item 4 History	-0.027
Item 5 History	0.123 *
Item 6 History	0.081
Item 7 History	0.073
Item 8 History	0.009
Item 9 History	0.051
Item 10 History	0.070
Item 11 History	0.026
Item 12 History	0.025
Item 13 History	-0.030
Item 14 History	0.020
Item 15 History	0.020
Item 16 Physical Exam	0.098
Item 17 Physical Exam	0.021
Item 18 Physical Exam	0.071
Item 19 Physical Exam	0.134 *
Item 20 Physical Exam	0.049
Item 21 Physical Exam	-0.091
Item 22 Physical Exam	0.023
Item 23 Physical Exam	-0.002
Item 24 Impressions	0.240 **
Item 25 Impressions	0.319 **
Write-up Legibility	0.038
Write-up History	0.316 **
Write-up Physical Exam	0.253 **
Write-up Diagnosis	0.384 **
Write-up Follow-up Plan	0.425 **
* Correlation is significant at the $p < 0.05$ level	
** Correlation is significant at the $p < 0.01$ level	

Appendix F: Case B Correlations between Items and Overall Rating

CASE B	
Item	**Correlation**
Item 1 Introductory	0.035
Item 2 Introductory	0.019
Item 3 History	n/a
Item 4 History	-0.030
Item 5 History	0.096
Item 6 History	0.068
Item 7 History	0.136 *
Item 8 History	0.128 *
Item 9 History	0.141 **
Item 10 History	0.264 **
Item 11 History	0.175 **
Item 12 History	0.199 **
Item 13 History	-0.006
Item 14 History	0.030
Item 15 History	-0.048
Item 16 Review of Systems	0.106 *
Item 17 Impressions	0.202 **
Item 18 Impressions	0.243 **
Write-up Legibility	0.173 **
Write-up History	0.450 **
Write-up Diagnosis	0.625 **
Write-up Follow-up Plan	0.421 **
* Correlation is significant at the $p < 0.05$ level ** Correlation is significant at the $p < 0.01$ level	

Appendix G: Case C Correlations between Items and Overall Rating

CASE C	
Item	**Correlation**
Item 1 Introductory	0.061
Item 2 Introductory	0.027
Item 3 History	0.147 **
Item 4 History	0.066
Item 5 History	0.109 *
Item 6 History	0.178 **
Item 7 History	0.203 **
Item 8 History	0.235 **
Item 9 History	0.220 **
Item 10 History	0.236 **
Item 11 History	0.255 **
Item 12 History	0.124 *
Item 13 History	0.059
Item 14 History	0.198 **
Item 15 Physical Exam	0.069
Item 16 Physical Exam	0.100
Item 17 Physical Exam	0.178 **
Item 18 Physical Exam	0.215 **
Item 19 Physical Exam	0.153 **
Item 20 Physical Exam	0.161 **
Item 21 Physical Exam	0.196 **
Item 22 Physical Exam	0.087
Item 23 Physical Exam	0.189 **
Item 24 Physical Exam	0.144 **
Item 24 Impressions	0.485 **
Item 25 Impressions	0.374 **
* Correlation is significant at the $p < 0.05$ level ** Correlation is significant at the $p < 0.01$ level	

Appendix H: Case D Correlations between Items and Overall Rating

CASE D	
Item	**Correlation**
Item 1 History	0.118 *
Item 2 History	0.265 **
Item 3 History	0.183 **
Item 4 History	0.366 **
Item 5 Diagnosis	0.266 **
Item 6 Diagnosis	0.285 **
Item 7 Diagnosis	0.245 **
Item 8 Ethics	0.124 *
Item 9 Ethics	0.118 *
Item 10 Ethics	0.087
Item 11 Ethics	0.266 **
* Correlation is significant at the $p < 0.05$ level ** Correlation is significant at the $p < 0.01$ level	

Appendix I: Correlations and Covariances between Revised Subscales

		CASE A						
		General History	Specific History	Physical Exam	Write-up Legibility	Write-up History	Write-up Physical Exam	Write-up Diagnosis
CASE A								
General History	Corr.	1						
	Covar.	0.316						
Specific History	Corr.	0.145	1.000					
	Covar.	0.128	2.460					
Physical Exam	Corr.	0.041	0.216	1.000				
	Covar.	0.038	0.563	2.754				
Write-up Legibility	Corr.	-0.054	-0.155	0.003	1.000			
	Covar.	-0.017	-0.132	0.003	0.293			
Write-up History	Corr.	0.157	0.244	0.132	-0.119	1.000		
	Covar.	0.047	0.203	0.117	-0.034	0.282		
Write-up Physical Exam	Corr.	0.112	-0.100	0.137	0.003	0.050	1.000	
	Covar.	0.037	-0.092	0.133	0.001	0.015	0.339	
Write-up Diagnosis	Corr.	0.085	-0.072	-0.042	-0.036	0.037	0.093	1.000
	Covar.	0.026	-0.062	-0.039	-0.011	0.011	0.030	0.307
Write-up Follow-up Plan	Corr.	0.054	-0.030	-0.068	0.003	0.124	-0.030	0.137
	Covar.	0.014	-0.021	-0.052	0.001	0.030	-0.008	0.034
CASE B								
History	Corr.	0.015	0.148	0.105	-0.088	0.131	0.086	0.066
	Covar.	0.026	0.712	0.534	-0.144	0.213	0.154	0.111
Write-up Legibility	Corr.	0.020	0.043	0.044	0.247	0.085	-0.037	0.066
	Covar.	0.009	0.053	0.057	0.105	0.036	-0.017	0.029
Write-up History	Corr.	0.126	0.129	0.159	0.008	0.322	0.085	0.038
	Covar.	0.069	0.196	0.255	0.004	0.166	0.048	0.021
Write-up Diagnosis	Corr.	-0.058	-0.033	-0.009	-0.008	0.016	0.019	-0.044
	Covar.	-0.032	-0.050	-0.015	-0.004	0.008	0.011	-0.024
Write-up Follow-up Plan	Corr.	0.067	0.030	-0.060	0.053	0.049	-0.062	0.046
	Covar.	0.039	0.049	-0.104	0.029	0.027	-0.038	0.027
CASE C								
History	Corr.	0.035	0.186	0.168	-0.028	0.208	-0.025	0.022
	Covar.	0.034	0.502	0.479	-0.027	0.191	-0.025	0.021
Physical Exam	Corr.	0.035	0.027	0.064	-0.029	0.032	0.034	0.024
	Covar.	0.044	0.093	0.234	-0.035	0.037	0.044	0.030
CASE D								
History/Diagnosis	Corr.	0.077	0.039	0.001	-0.115	0.068	0.013	0.070
	Covar.	0.063	0.089	0.002	-0.091	0.053	0.011	0.056
Ethics	Corr.	0.018	0.141	0.155	0.058	0.076	0.022	0.055
	Covar.	0.008	0.173	0.200	0.025	0.031	0.010	0.024

Appendix I: Correlations and Covariances between Revised Subscales (Cont.)

		CASE A	CASE B				
		Write-up Follow-up Plan	History	Write-up Legibility	Write-up History	Write-up Diagnosis	Write-up Follow-up Plan
CASE A (cont.)							
Write-up Follow-up Plan	Corr.	1.000					
	Covar.	0.209					
CASE B							
History	Corr.	-0.006	1.000				
	Covar.	-0.008	9.371				
Write-up Legibility	Corr.	0.049	0.078	1.000			
	Covar.	0.018	0.188	0.623			
Write-up History	Corr.	0.036	0.313	0.109	1.000		
	Covar.	0.016	0.927	0.083	0.936		
Write-up Diagnosis	Corr.	0.006	0.195	0.120	0.209	1.000	
	Covar.	0.003	0.589	0.093	0.200	0.974	
Write-up Follow-up Plan	Corr.	0.159	0.043	0.168	0.027	0.215	1.000
	Covar.	0.075	0.136	0.138	0.027	0.221	1.083
CASE C							
History	Corr.	0.095	0.235	0.103	0.232	0.083	0.062
	Covar.	0.075	1.239	0.141	0.388	0.142	0.112
Physical Exam	Corr.	-0.052	0.033	0.035	0.015	0.027	0.018
	Covar.	-0.053	0.222	0.061	0.032	0.058	0.042
CASE D							
History/Diagnosis	Corr.	0.017	-0.011	0.225	0.178	0.108	0.189
	Covar.	0.011	-0.048	0.259	0.251	0.155	0.287
Ethics	Corr.	0.127	0.032	0.116	0.063	0.009	0.080
	Covar.	0.045	0.077	0.072	0.048	0.007	0.065

		CASE C		CASE D	
		History	Physical Exam	History/ Diagnosis	Ethics
CASE C					
History	Corr.	1.000			
	Covar.	2.974			
Physical Exam	Corr.	0.072	1.000		
	Covar.	0.276	4.863		
CASE D					
History/Diagnosis	Corr.	0.131	0.071	1.000	
	Covar.	0.331	0.228	2.126	
Ethics	Corr.	0.070	0.039	0.108	1.000
	Covar.	0.095	0.068	0.123	0.609

Appendix J: Correlations between Subscales and USMLE Step 2 CK Scores

New Subscale	Correlation	
Case A History General	0.048	
Case A History Specific	0.050	
Case A Physical Exam	0.045	
Case A Write-up Legibility	0.036	
Case A Write-up History	0.071	
Case A Write-up Physical Exam	0.163	**
Case A Write-up Diagnosis	0.014	
Case A Write-up Follow-up Plan	0.042	
Case B History	0.037	
Case B Write-up Legibility	0.040	
Case B Write-up History	0.092	
Case B Write-up Diagnosis	0.041	
Case B Write-up Follow-up Plan	0.119	*
Case C History	0.049	
Case C Physical Exam	0.186	**
Case D History and Diagnosis	0.115	*
Case D Ethics	0.070	

* Correlation is significant at the $p < 0.05$ level
** Correlation is significant at the $p < 0.01$ level

VITA

Kirsten Roberts was born in Tripoli, Libya and grew up living in many places throughout the world. Currently she calls Seattle home. At Texas A&M University she earned a Bachelor of Business Administration and at Texas State University she earned a Master of Business Administration. In 2007 she earned a Doctor of Philosophy at the University of Washington in Educational Psychology.