UNIVERSITY OF CALGARY

Reliability, validity and sources of errors in assessing physician performance in an Objective Structured Clinical Examination: A Generalizability Theory Analysis.

by

Andrea Laurie Cameron Vallevand

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF DOCTOR OF PHILOSOPHY

FACULTY OF KINESIOLOGY

CALGARY, ALBERTA

APRIL, 2008

UNIVERSITY OF CALGARY

FACULTY OF GRADUATE STUDIES

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies for acceptance, a thesis entitled "Reliability, validity and sources of errors in assessing physician performance in an Objective Structured Clinical Examination: A Generalizability Theory Analysis." submitted by Andrea Laurie Cameron Vallevand in partial fulfilment of the requirements of the degree of Doctor of Philosophy.

Supervisor, Dr. David Paskevich, Faculty of Kinesiology

Claudio Violato

Co-Supervisor, Dr. Claudio Violato, Faculty of Medicine

Dr. Tina Gabriele, Faculty of Kinesiology

Jocelyn Lockyer

Dr. Jocelyn Lockyer, Faculty of Medicine

Dr. Tanya Beran, Faculty of Education

External Examiner, Dr. Richard Mrazek,
University of Lethbridge

Feb 20, 2008

Date

Abstract

Background: International Medical Graduates (IMGs) from British Columbia, Alberta, Saskatchewan, Manitoba, and the Yukon participated in a high-stakes, fourteen case Objective Structured Clinical Examination (OSCE). The OSCE was part of the clinical competency assessment process developed for a demonstration project funded by Health Canada to incorporate qualified IMGs into the Western Canadian physician workforce.

Purpose: The primary purpose of the present study was to identify the sources of error affecting the reliability of assessments of physician competence. The secondary purpose was to evaluate the reliability of physician versus non-physician examiners to score performance-based assessments from videotape recordings of OSCE cases. Generalizability analyses and other statistical procedures were utilized to study the psychometric properties of an OSCE developed to evaluate clinical performance in several domains such as history taking, application of physical examination skills, and the ability to communicate and/or counsel patients.

Methods: Thirty-nine candidates rotated through fourteen 12-minute cases. Ten minutes were assigned for the candidate/standardized patient (SP) interaction, which was followed by a two-minute post-encounter probe. Each case was evaluated using a case-specific checklist, a five point global rating of overall performance, and a communication skills checklist. One presiding physician examiner completed all three of the assessment instruments, while the SP only completed the communication skills checklist. For the physician versus non-physician study, the physician rater trained two non-physician raters on the necessary assessment protocols for each case. Fifteen candidates were randomly selected and their videotaped interactions with the SP were evaluated using the case-specific checklist and communication skills checklist.

Results: The reliability coefficients ranged from low to moderate on the assessment formats. Generalizability analyses revealed large variance components attributed to SPs and examiners which suggest the candidates' scores were being unduly influence by physician examiners and SPs. The communication skills assessment revealed moderate to high internal consistency for each case. Low to moderate internal consistency was calculated for each checklist item across cases (inter-item). The percent variance for each checklist on participants, cases, p x c, and for the nested SPs and assigned physician examiners varies considerably from item to item, suggesting are differences in how the items are being rated based on the requirements of the case.

Physician raters and non-physician raters showed statistically significant differences on the checklist score in four cases, differences in the global scores on three cases, and differences in the total score on three cases. The intra-class coefficient ranged low to high on the physician examiner data and moderate to high on the non-physician examination data.

Conclusions: Low reliability results and large variance components attributed to physician examiners and SPs call into question the reliability of the OSCE for high-stakes assessment. Standardized training by the videotape physician rater aided in higher reliability coefficients with the checklist scores but the coefficients were lower on the global scores.

Acknowledgements

I am going to make this short; which will likely surprise everybody I know.

I would like to begin by thanking my supervisors, Dr. Dave Paskevich and Dr. Claudio Violato. My sincere appreciation to you both for your knowledge and support.

I would like to thank the other members of my committee for their efforts: Dr. Tanya Beran, Dr. Tina Gabriele, Dr. Jocelyn Lockyer, and Dr. Rick Mrazek. Thank you, also, to Dr. David Cawthorne for being a member of my Candidacy committee.

A big ‘shout out’ (and my eternal gratitude) goes to Dr. Gail Kopp and Dr. Margo Mayo for their belief and encouragement. Having wonderful people you trust (with your life) watching your back makes all things seem possible.

Table of Contents

CHAPTER FOUR

CHAPTER FIVE

List of Tables

List of Figures

CHAPTER ONE

INTRODUCTION

When human assessors are involved in the evaluation of performance, errors of measurement are commonplace (Downing, 2004). This is true for evaluating athletic skills, cognitive performance and medical competence. Such errors can have an important impact on the reliable and valid assessment of performers. From a medicine standpoint, the assessment of competency in medical education and physician licensing is essential to public accountability (Shumway & Harden, 2003). Regardless of whether the assessment is for formative feedback to medical students and residents (Carraccio & Englander, 2000; Crossley, Humphris, & Jolly, 2002; Harden, Stevenson, Wilson Downie, & Wilson, 1975; Miller, 1990), for summative evaluation at the end of a course or program (van der Vleuten, 2000), or for ensuring a minimum level of clinical competence (Crossley, Humphris et al., 2002), the assessment process must be consistent, accurate and defensible (Boulet, McKinley, Whelan, & Hambleton, 2003).

Performance-based assessments, for example the Objective Structured Clinical Examination (OSCE), are prone to a large range of potential measurement errors (Boulet et al., 2003). These sources of error can be categorized as error due to examination content and error due to scoring inconsistencies or mistakes. Measurement errors can be minimized by ensuring that case development covers the necessary content and skills and that there is consistent training of the standardized patients (SPs) and examiners. Upon completing the examination process, Boulet and colleagues (2003) recommended that generalizability analyses be used to identify any sources of error in a particular examination. Examination content error and errors of measurement will influence the reliability of the evaluation and subsequently the validity of the assessment.

Performance-Based Assessment (OSCE)

OSCEs are primarily suited for assessing clinical, technical and practical skills (Newble, 2004). During a typical OSCE, the candidates rotate through a series of five to ten minute cases that have been designed to evaluate performance on specific tasks (e.g., history taking or interpreting a diagnostic test) (Hilliard & Tallett, 1998; Newble, 2004; Wass, van der Vleuten, Shatzer, & Jones, 2001). Candidates are generally assessed with a case-specific checklist brandished by an examiner or examiners (Hilliard & Tallett, 1998; Newble, 2004; Petrusa, Blackwell, & Ainsworth, 1990). The purpose of the checklist is to increase the reliability of the assessment by identifying the distinct parameters of what comprises unacceptable and acceptable performance (Winckel, Reznick, Cohen, & Taylor, 1994). Candidates might also be evaluated with global rating scale(s), which can be utilized to assess the fluency of skill application or evaluate the candidates' overall clinical performance (Norman, van der Vleuten, & De Graaff, 1991).

As part of the assessment process, an OSCE might incorporate real patients who have volunteered to participate in candidate training or evaluation (Adamo, 2003; Sloan, Donnelly, Schwartz, Vasconez, Plymale, & Kenady, 1998; van der Vleuten & Swanson, 1990). Generally though, OSCEs have trained actors or non-actor volunteers portraying the role of the patient (Adamo, 2003). The term 'standardized patient' (SP) indicates that the content of the responses (verbal and behavioral) are presented in a uniformed and consistent manner (Adamo, 2003; van der Vleuten & Swanson, 1990).

Trained, experienced physicians are considered the most qualified raters of candidate performance on performance-based assessments (Humphrey-Murto, Smee, Touchie, Wood, & Blackmore, 2005; Martin, Reznick, Rothman, Tamblyn, & Regehr, 1996). When factors such as cost (Boulet et al., 2003; Martin et al., 1996; McLaughlin, Gregor, Jones, & Coderre, 2006) and availability of physician raters (e.g., recruitment) (Boulet et al., 2003; Humphrey-Murto et al., 2005; McLaughlin et al., 2006) are issues, the decision as to whether trained non-physician raters can be used to reliably evaluate candidate performance must be carefully considered.

Project Background

The Western Alliance for Assessment of International Physicians (WAAIP) was a demonstration project funded by Health Canada to incorporate qualified International Medical Graduates (IMGs) into the Western Canadian (British Columbia, Alberta, Saskatchewan, Manitoba) and Northern Canadian (Yukon, Northwest Territories, Nunavut) physician workforce (Western Alliance for Assessment of International Physicians, 2006). IMGs are physicians who have graduated from a medical school outside of Canada or the United States (Crutcher, Banner, Szafran, & Watanabe, 2003).

While some IMGs are able to find professional positions in under-served areas (e.g. rural communities), many are landed immigrants or refugees who are unable to gain employment as they require post-graduate training within a residency program in order to become licensed to practice in Canada (Andrew & Bates, 2000). The WAAIP project was designed to provide an alternate route into medical practice for IMG's whose knowledge and skills were sufficiently advanced that the completion of a full residency program would not be required (Violato & Baig, 2006).

Potential candidates underwent a rigorous selection process to qualify for the WAAIP project. The successful candidates must have graduated from a medical school included in the World Health Organization's directory of medical institutions and had his or her medical degree verified by the Educational Commission for Foreign Medical Graduates International Credentials Services. Each candidate must have met the minimum required standards on the Test of English as Foreign Language (TOEFL) and have passed the Medical Council of Canada Evaluating Examination. From an initial pool of one hundred and sixteen candidates; 39 were selected to participate (Violato & Baig, 2006).

WAAIP OSCE Procedures

The candidates participated in a fourteen case OSCE administered at the Medical Skills Centre, Faculty of Medicine at the University of Calgary. Each case was 12 minutes in duration with the candidate/SP interaction scheduled for ten minutes and the post-encounter probe encompassing the last two minutes. The post-encounter probe was conducted by the physician examiner and consisted of a case-specific series of questions (e.g., what are your differential diagnoses?).

Prior to entering a testing room, the candidate was given three minutes to review a one page synopsis of the case (e.g., patient complains of a fever) and the case's requirements (e.g., take a patient history and perform a focused physical examination). Each case was evaluated using a case-specific checklist, one five-point global rating of overall performance, and a communication skills checklist. One presiding physician examiner completed all three of the assessment instruments and the assigned SP completed the communication skills checklist. All of the candidates' performances were videotaped using an unobtrusive, built-in audio and video recording system. For the purposes of a secondary statistical evaluation, randomly selected videotaped performances were scored by one physician and two non-physician raters.

Statement of the Problem

Identified sources of error in performance-based evaluations include inappropriately designed cases (Boulet et al., 2003; Petrusa, 2002), poorly trained examiners (Downing & Haladyn, 2004), and inadequately trained SPs (Boulet et al., 2003). From a case design standpoint, thorough case development is one the most effective strategies for reducing measurement error as reliability will be affected if the case measures the skills inadequately (Boulet et al., 2003). Downing (2003b) specified several principles for guiding case design such as the cases must be representative of the course/rotation blueprint and evidence must be presented that faculty content experts have designed, reviewed and revised the cases.

Examiner inconsistency or low inter-rater reproducibility has been identified as the largest threat to the reliability of an assessment (Downing, 2004). Standardized training of the examiners to use the assessment instruments is critical (Boulet et al., 2003; van der Vleuten & Swanson, 1990; Wass et al., 2001).

Regarding SPs, measurement error can be introduced via the inconsistency of SP performance (Boulet et al., 2003; van der Vleuten & Swanson, 1990), inaccurate performance by the SP, the choice of SP to portray the case, and/or the SP's portrayal of the case (Boulet et al., 2003). Accuracy of SP portrayal is critical during performance assessments when the candidate is gathering information to formulate a diagnosis (or differential diagnoses) and creating the subsequent treatment and management plans

(Tamblyn, Klass, Schnabl, & Kopelow, 1991). Reproducibility of SP portrayal is fundamental to the claim that the same case is presented to every candidate (Petrusa, 2002). Properly trained SPs do not vary in their presentation from candidate to candidate (Barrows, 1987). SPs that are not trained to consistently portray the patient in a standardized manner will likely result in different examinees effectively encountering different patients and slightly different patient problems (Downing, 2003b).

While Boulet and colleagues (2003) have emphasized the potential for a variety of SP-related measurement error, Williams (2004) suggested that refinements in SP training methods have likely led to improvements in the accuracy of SP portrayal, but was unaware of any studies subsequent to Tamblyn (1989) that have evaluated SP accuracy and reproducibility and its potential impact on student performance outcomes.

Tamblyn (1989) assessed SP accuracy within the framework of an end of the year OSCE involving final year medical students. The first study focused on the accuracy of SP performance while the second study concentrated on the impact of SP accuracy on the medical students' performance scores. The results of the first study indicated that a 90% or higher accuracy was achieved by almost two-thirds of the SPs trained for the examination. Of the errors identified, one third were categorized as systemic (the error consistently occurs, which indicates there is a training problem) and two-thirds were random error (an incorrect response is provided on some but not all of the SP/candidate interactions). While the results of the second study indicated no relationship between SP accuracy and the overall clinical competence score, there was one case where the two SPs' accuracy scores were high (100% and 98%) but an error in the presentation of one item in the case had consequences for diagnosis and case management.

Adamo (2003) stated that standards and models for SP training have generally not been documented. While standards and models do exist in practice, they vary from institution to institution. Recruitment, thorough training and the application of quality assurance protocols are essential for the continued accurate performance of SPs for clinical evaluation (Adamo, 2003; Glassman, Luck, O'Gara, & Peabody, 2000), although, Adamo (2003) further observed there are often few resources available to conduct quality assurance reviews of SP programs.

Boulet and colleagues (2003) have recommended several quality assurance methods to evaluate whether the scores collected are reliable and relatively free from measurement error. These strategies include internal consistency measures, item analysis, correlation analyses between individual case scores and the mean scores across the remaining cases, and generalizability analyses. While some recent studies have used generalizability analyses in their statistical analyses (Guiton, Hodgson, Delandshere, & Wilkerson, 2004; Murray, Boulet, Kras, McAllister, & Cox, 2005; Weller, Robinson, Jolly, Watterson, Joseph, Bajenov et al., 2005), the author of this study has been unable to find studies that have analyzed the variance components for SPs and SP/rater combinations, nor have studies been found that utilized the extensive quality assurance assessment protocols that were outlined by Boulet and colleagues (2003).

Purpose of the Present Study

There are many potential sources of error and threats to reliability in clinical competency assessment. The primary purpose of the present study was to identify any sources of error affecting the reliability of assessments of physician competence and performance. Generalizability analyses and other statistical procedures were utilized to study the psychometric properties of the WAAIP OSCE; an examination which was developed to evaluate the clinical performance of IMGs in several domains such as history taking, application of physical examination skills, and the ability to communicate with and/or counsel patients. The secondary purpose was to evaluate the reliability of physician versus non-physician raters to score performance-based assessments from videotape recordings of OSCE cases.

If the reliability of this fourteen case OSCE can be established, the results could provide a model for identifying the characteristics, strengths and limitations of performance-based assessment examinations designed to evaluate the practice readiness of medical graduates. Moreover, this may have application to assessment of human performance in other domains (e.g., athletic performance and psychomotor skill assessments).

Overview of the Dissertation

Chapter one provided an overview of the study. Chapter two is designed to familiarize the reader with the concepts of OSCEs, SPs, the psychometrics of performance-based assessments, assessment formats, and generalizability analysis. Chapter three is the methods section and presents the candidates, outlines the format of the fourteen case OSCE and its particulars, lists the statistical analyses to be used for both study one and study two. Chapter four presents the results of both studies, and Chapter five presents the discussion and conclusions of this dissertation.

CHAPTER TWO

REVIEW OF LITERATURE

Chapter two will focus on the following seven topics: the Objective Structured Clinical Examination, standardized patients, the psychometrics of performance-based assessments, assessment formats in performance-based assessments, physician versus non-physician evaluation, sources of error in performance-based assessment, and generalizability analysis. [1]

Objective Structured Clinical Examination

Introduction

Initially called the 'structured clinical examination' (Harden et al., 1975), the subsequently named Objective Structured Clinical Examination (OSCE) (Harden & Gleeson, 1979) was designed to be a controlled assessment of clinical competence by removing the variability introduced by the patient and the examiner (Harden & Gleeson, 1979; Harden et al., 1975). The OSCE was established as a tool for medical student assessment by Harden and colleagues in the mid 1970's (Harden et al., 1975). It was designed to be a practical, valid, and reliable evaluation strategy to control for examiner biases found in other student evaluation methods, for example, the traditional clinical examination (Carraccio & Englander, 2000; Harden & Gleeson, 1979). In these

[1] The present dissertation is focused on the psychometric evaluation of a fourteen case Objective Structured Clinical Examination. For those interested in the International Medical Graduate literature, the following resources are recommended: Dauphinee (2005), Crutcher, Banner, Szafran and Watanabe (2003), Andrew and Bates (2000), Muller, Harik, Margolis, Clauser, McKinley, and Boulet (2003), Norcini and Mazmanian (2005).

traditional clinical examinations, a student's competence was assessed using a limited number of patients found in the hospital ward (Harden et al., 1975). Problems included the presentation of the patient; often the cases were not commonly encountered in clinical practice, typically the cases were chronic versus acute, and certain specialties (e.g., otolaryngology) were not or under represented (Harden & Gleeson, 1979). Furthermore, the scores awarded often varied considerably between examiners (Harden & Gleeson, 1979; Harden et al., 1975).

Advantages and Disadvantages of the OSCE

Harden et al. (1975) and Harden and Gleeson (1979) identified several advantages of the OSCE. These included the ability to control the complexity of the case for varying skill levels of students, clearly defining the knowledge, skills, and attitudes to be assessed, and creating an examination that could sample a wider range of knowledge and skills including those not often seen in the traditional clinical examination (e.g., management of an emergency situation). In addition to measuring clinical competence while controlling for observer biases, an OSCE can be specifically designed for the formative assessment of students (Hilliard & Tallett, 1998), which can subsequently be used to identify students who are performing at less than acceptable levels (Hilliard & Tallett, 1998; Martin & Jolly, 2002). Harden et al. (1975) and Harden and Gleeson (1979) also noted that the use of an objectively evaluated examination allowed for the comparison of standards across cohorts of students (e.g., second year medical students over a multi-year period) and provided an opportunity for structured feedback of both students and faculty.

Disadvantages include the extensive faculty time commitment required (Carraccio & Englander, 2000) and the expense in terms of resources needed (e.g., testing rooms) and personnel required (e.g., standardized patients) (Carraccio & Englander, 2000; Mavis & Henry, 2002; Wass et al., 2001). There have also been concerns expressed that students might compartmentalize knowledge instead of looking at the patient as a whole (Harden & Gleeson, 1979; Harden et al., 1975).

The OSCE and Clinical Competence Assessment

Within the framework of Miller's Pyramid of Competence (Figure 1) (Miller, 1990) the OSCE fits within the 'Shows how' category which reflects the ability of the participant to demonstrate behaviours within a practice or simulated situation (van der Vleuten, 2000; Wass et al., 2001).

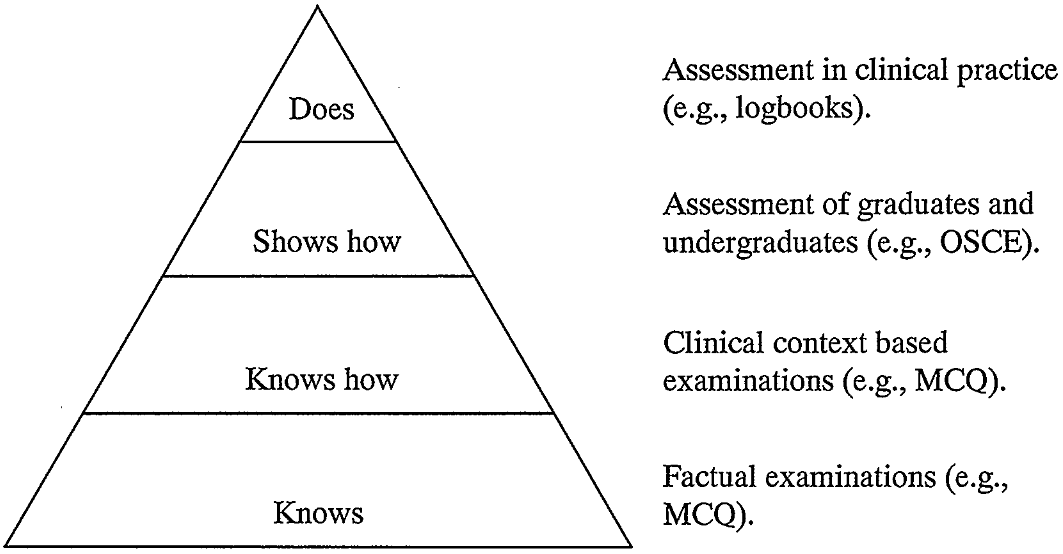
Does
Shows how
Knows how
Knows
Assessment in clinical practice (e.g., logbooks).
Assessment of graduates and undergraduates (e.g., OSCE).
Clinical context based examinations (e.g., MCQ).
Factual examinations (e.g., MCQ).

Figure 1. Miller's pyramid of competence (Miller, 1990).

The OSCE is principally suited for assessing technical, clinical and practical skills with the typical being OSCE comprised of a series of five to ten minute cases (Newble, 2004) or depending on the task to be performed; five to twenty minutes (van der Vleuten & Swanson, 1990). Longer cases can be utilized depending upon the skill level of the students and the required authenticity of the clinical situation (e.g., a psychiatric assessment) (Hodges, 2003). During an OSCE, the candidates rotate through a series of cases designed to evaluate performance on specific clinical tasks (e.g., history taking or interpreting a diagnostic test) (Hilliard & Tallett, 1998; Wass et al., 2001). The candidates can be evaluated with a objective criteria checklist and/or a global rating scale(s) wielded by an examiner or examiners (Hilliard & Tallett, 1998; Newble, 2004; Petrusa et al., 1990). The purpose of the objective criteria checklist is to increase the reliability of the assessment by identifying the distinct parameters of what comprises unacceptable and acceptable performance (Winckel et al., 1994). Global scales are often used to evaluate the more difficult to define skills (e.g., the fluency of the candidate's physical assessment skills) or overall performance on a particular case (Norman et al., 1991; Streiner, 1985).

Cases will usually have trained actors or non-actor volunteers playing the role of the SP (Adamo, 2003) although on some occasions a case might incorporate a real patient who has volunteered to participate (Adamo, 2003; Sloan et al., 1998; van der Vleuten & Swanson, 1990). The term 'standardized' indicates that the content of the responses (from both behavioral and verbal standpoints) are presented in a standardized and consistent manner (Adamo, 2003; van der Vleuten & Swanson, 1990).

Standardized Patients

Introduction

The use of SPs was introduced by Barrows in the early 1960's (Adamo, 2003; Barrows, 1993). Originally called 'programmed patients' and later 'patient simulators' (Barrows, 1971), SPs were used in the classroom and non-clinical settings to present the same clinical problem to a series of medical students (Barrows, 1993). The SP provided a transition to real patients in an effort to give beginning medical students an opportunity to become confident in their history taking and physical assessment skills and be exposed to cases they normally would not be allowed to manage (e.g., sensitive circumstances or emergency situations) (Barrows, 1987; Barrows, 1993). SPs are available at any time, are available in both clinic and non-clinic settings, allow for interaction with 'difficult' patients, present opportunities for feedback on performance, and afford the novice time to work with case or allow time constraints to be provided to the more experienced student (Barrows, 1987).

Training SPs

Training begins with the selection of a clinical case. Often cases are based on real patients in case files. Fictional cases are created when complexities need to be removed or key desirable factors added. Controlling the complexity of a case is generally considered appropriate for a case that is presented early in medical school. Once the patient's case has been determined, the script is created in a clear, detailed, and 'sans medical terminology' manner. Attention to detail is critical. Items include: signs and symptoms, details on the onset of the illness, important background information (e.g., past pertinent medical history), and family history. Barrows (1971) recommended a physician be used to train/coach the SP. The coaching must continue until the SP's performance matches the physician's perception of how a real patient would look and act. In addition, others who have cared for a similarly ill patient should be involved in the training process.

The next step focuses on the complete picture of the illness from the patient's point of view (taking on the illness as 'his or her own'). For example, how the SP would have noticed the onset of the symptoms or what would be the emotional response to the onset of the illness. The SP needs to be trained on what information should be volunteered to the medical student and what information should be presented only when the medical student asks. Education about the illness and the use of medical terminology must be strictly avoided.

The training of physical signs can be intensive and repetitive. Throughout the process the physician/coach must assume an active role and continually work with the SP. Trial performances (rehearsals) with feedback are mandatory. Many ingenious strategies have been created for training SPs to portray physical signs of illness or injury (Barrows, 1987; Barrows, 1999) and well trained SPs can be very realistic to the point where the medical students forget they are working with simulated patients and relate to simulated patients as though they are a real patients (Barrows, 1971; Barrows, 1987). Finally, an independent clinical physician should 'work up' the SP case at the completion of training and provide feedback regarding the believability of the SP's performance.

The Psychometrics of Performance-Based Assessment

The purpose of any assessment protocol is to provide inferences about the ability or competency of the candidates – inferences that extend beyond the sample of cases or cases included in the examination (Swanson, 1987; van der Vleuten & Swanson, 1990). Regardless of whether the assessment is used for formative feedback to medical students and residents (Carraccio & Englander, 2000; Crossley, Humphris, & Jolly, 2002; Harden et al., 1975; Miller, 1990), designed for summative evaluation at the end of a course or program (van der Vleuten, 2000), or utilized to ensure a minimum level of competence (Crossley, Humphris et al., 2002) the assessment process must be consistent, accurate and defensible (Boulet et al., 2003).

Reliability

Reliability refers to the consistency or reproducibility of test scores (Downing, 2004; Shea & Fortna, 2002; Streiner & Norman, 1989; van der Vleuten, 2000; Wass et al., 2001) or to the precision of the measurement (van der Vleuten, 2000). Reliability is a major source of validity evidence for all assessments (Downing, 2004). Unless the evidence collected is reliable it becomes almost impossible to interpret whether the assessment is valid (Downing, 2003b).

Shea and Fortna (2002) identified two types of reliability: internal consistency (e.g., Cronbach's alpha) and reproducibility (e.g., inter-rater reliability). A third type of reliability is temporal stability which is typically assessed as test-retest reliability (Violato, Marini, & McDougall, 1998). Internal consistency measures whether items on an examination measure the same construct (e.g., history taking) and it provides a summary as to how well a set of items measure the same general construct. (Cronbach et al., 1972; Shea & Fortna, 2002; Streiner & Norman, 1989). Reproducibility (or repeatability) refers to whether scores collected on one occasion are the same as scores collected on another. An example is inter-rater reliability, which refers to the agreement between two or more raters on an assessment (Shea & Fortna, 2002). The preferred statistical procedure for calculating reproducibility is an intra-class correlation (ICC) coefficient (Shea & Fortna, 2002), which is based on the analysis of variance (Cronbach et al., 1972; Shea & Fortna, 2002; Streiner & Norman, 1989).

A reliability coefficient of at least 0.90 is recommended for very high stakes assessment (certification or licensure) (Downing, 2004; Shea & Fortna, 2002). For moderate stakes assessments (end of course or end of year summative assessment) the reliability should range between 0.80 to 0.89. The reliability for classroom summative and formative assessments should range between 0.70 to0.79 (Downing, 2004), while reliability for educational research can range from 0.60 to 0.80 (Shea and Fortna, 2002).

Threats to reliability. An assessment with a small number cases can produce unstable or unreliable scores (Downing, 2004; van der Vleuten, 1996; van der Vleuten & Swanson, 1990). Several cases are required to ensure that there is wide enough sampling to maintain an acceptable degree of reliability (Carraccio & Englander, 2000; Downing, 2004; Newble, 2004; Strube, 2000; Wass et al., 2001). Another threat to reliability is not having enough cases to evaluate a specific construct (e.g., physical assessment skills) (Boulet et al., 2003; Newble, 2004). Reliability of the assessment can also be affected if the case measures the skill(s) inadequately or if the assessment items are too specific for the content of the case (Boulet et al., 2003). Competence is content specific and competence in one case (e.g., the patient's chief complaint is chest pain) is not predictive of competence in another (e.g., where the patient's chief complaint is abdominal pain) even if the cases are closely related (Crossley, Humphris et al., 2002). For this reason, a wide sampling of topics is necessary for the evaluation of clinical competence (van der Vleuten, 2000; Wass et al., 2001).

Strategies to Increase Reliability. van der Vleuten (1996) stated that in order for any measurement of performance to be reliable there must be a sufficient sample of observations and these observations should be gathered with some degree of structure and standardization. Standardization refers to the controlling of measurement conditions in order to reduce the amount of error that might influence the scores. Aggregation refers to increasing the number of items on a test. In this way, the random sources of error that can influence the observed scores have an opportunity to cancel out; allowing for a better estimate of the true score (Strube, 2000).

Standardized patients, case design, and examiners can all influence the reliability of the observed scores (Boulet et al., 2003; Downing & Haladyn, 2004; van der Vleuten, 1996). Detailed training of SPs is required; especially when there are multiple SPs trained for a particular case (Adamo, 2003; Boulet et al., 2003; Carraccio & Englander, 2000; Downing & Haladyn, 2004; van der Vleuten & Swanson, 1990). Meticulous case design is one the most efficient strategies for reducing measurement error in performance-based examinations (Boulet et al., 2003), and the standardized training of examiners on the use of the assessment instruments (e.g., global rating scale) is critical (Boulet et al., 2003; Wass et al., 2001).

Assessment Formats in Performance-Based Assessment

Early clinical skills development might best be evaluated using a checklist format. When more difficult to quantify elements are being evaluated in more advanced students (e.g., the organization or fluency of the physical assessment), checklists might not capture these dimensions and thus render the evaluation less valid. Future research on the most appropriate assessment format for the skill level of the performer is essential; as is research on the potential limitations of trained laypersons to evaluate higher orders of clinical performance such as the implicit logic behind the evaluation (Petrusa, 2002).

Checklist format

van der Vleuten and Swanson (1990) and Norman et al. (1991) advocated the use of objective criteria checklists for isolated simple tasks performed by more novice medical students; as the checklist format can clearly outline what is expected of the candidate. Objective criteria checklists are also valuable for providing structured and specific feedback to the candidate (Norman et al., 1991; Petrusa, 2002; van der Vleuten & Swanson, 1990). There are some concerns, in the literature, regarding the use of the checklist format.

Crossley et al. (2002) suggested that is difficult to reduce the complexity of a clinical assessment into a checklist format. Norman et al. (1991), van der Vleuten and Swanson (1990) and Newble (2004) cautioned that the checklist format could result in the trivialization of knowledge. The use of checklists might also reward thoroughness (Cunnington, Neville, & Norman, 1997; Reznick, Regehr, Yee, Rothman, Blackmore, & Dauphinee, 1998; van der Vleuten, 1996) and penalize efficiency rather than evaluate the competence of the evaluation (Reznick et al., 1998) or whether there was a coherent approach in the gathering of information (Cohen, Rothman, Poldre, & Ross, 1991; Cunnington et al., 1997).

Newble (2004) suggested that the development of detailed checklists might result in reliable scores but not necessarily reflect performance when easy to define elements are included in the checklist (e.g., asks about the severity of pain [the physician might ask "on a scale of 1 to 10 with 10 being the most severe pain you have ever had – how would you rate the pain in your chest"]) while more difficult components to measure or characterize are not included (e.g., fluency of physical assessment skills). For this reason, the use of global rating scales have been recommended as an adjunct to address the limitations of the checklist format (Cohen et al., 1991).

Global rating format

Norman et al. (1991) recommended the use of global ratings to evaluate skills tested at more advanced levels of performance (e.g., residents). van der Vleuten and Swanson (1990) advised that difficult to articulate constructs, such as attitudes and communication skills, would also be better evaluated using global rating scales. van der Vleuten, Norman, and de Graffe (1991) and van der Vleuten (1996) proposed that global ratings may produce equally reliable assessments and evidence is mounting that a global ratings scored by a physician rater is as reliable as a checklist score (Wass et al., 2001). A limitation to global assessments is the extensive training of examiners that is required to ensure the consistency of scoring (Wass et al., 2001).

'Key Actions or Key Features' format.

A third assessment format focuses on the key actions or key features of the case (Murray et al., 2005; Page & Bordage, 1995). Page and Bordage (1995) developed the 'key features' assessment protocol founded on research that suggested that the effective assessment and management of a case is based on the manipulation of a few key components of that case. Williams (2004) and Page and Bordage (1995) advocated performance-based assessment using the key features format.

Physician versus Non-physician raters

van der Vleuten (1996) cautioned against the use of assessment protocols that avoided professional judgment (e.g., a physician rater). van der Vleuten and Swanson (1990) stated that physician examiners are more familiar with the logical sequencing of the history taking and physical assessment constructs of the examination, in addition to being able to better evaluate the technical proficiency of any applied physical assessment procedure(s). Reznick et al. (1998) echoed this caution stating that an expert examiner is relegated to the role of observer when assessment checklists are used for evaluation.

One consideration in the use of non-physician raters surrounds the type of assessment format to be used (Humphrey-Murto et al., 2005; Norman et al., 1991; van der Vleuten, Norman, & De Graaff, 1991). From a reliability standpoint, the research has demonstrated that well-trained SPs (non-physician raters) were able to accurately portray, recall and record the items addressed (e.g., history and physical assessment) during the examinee's clinical performance. Furthermore, SPs are able to evaluate the examinee's clinical performance as accurately as a physician rater (Colliver, Robbs, & Vu, 1991). The research has also indicated that that when SPs make errors in assessment, they typically err on the side of giving the examinee credit for an action that was not addressed (Vu, Barrows, Marcy, Verhulst, Colliver, & Travis, 1992). Williams recommended when

SPs are used to record examinee actions, especially in high-stakes examinations, a separate observer record the actions in real time to optimize accuracy.

van der Vleuten and Swanson (1990) stated that adequate inter-rater agreement can be achieved through the use of SPs or physicians. They cautioned, however, that both groups could vary in what aspects of examinee performance that can be accurately rated. They suggested that physician raters should be better able to evaluate whether the examinee demonstrated a logical sequencing of the evaluation and whether the physical assessment skills administered were performed in a technically proficient manner. SPs, on the other hand, should be better able to evaluate the examinee's communication skills.

Petrusa (2004), commenting on the use of SPs to evaluate performance, stated that clinically competent focused evaluations by senior level medical students must exhibit the logical and organized collection of data to demonstrate that the students can develop an array of differential diagnoses. In many cases, the SP simply records whether checklist items were addressed (or not) versus whether or not the candidate displayed a logical and organized evaluation. As a result, these important aspects of the candidates' performance (e.g., fluency in skill application) are not evaluated which is a serious oversight in the assessment of clinical performance. However, an SP simply records whether checklist items were addressed (or not), as opposed to whether the candidate displayed a logical and organized evaluation. As a result, these important aspects of the candidates' performance (e.g., fluency in skill application) are not evaluated which is a serious oversight in the assessment of clinical performance.

Utilizing both checklist and global assessments, Humphrey-Murto and colleagues (2005) investigated the amount of agreement between trained non-physician raters and physician raters on three of twelve OSCE cases that comprise the Medical Council of Canada Qualifying Examination Part II. The majority of the trained non-physician raters had some type of medical background (e.g., nurse or paramedic) while the remainder had previous experience as SPs. The findings indicated that there was overall good agreement between the physician and non-physician raters on the checklist scores (the main effect for examiner type was not statistically significant on all three cases). There was poor agreement on the global assessment (one six-point scale) where the physician raters failed more candidates than the non-physician raters on all three cases (14, 17, and 25 percent more, respectively). The authors concluded that trained non-physician raters did not possess the experience, knowledge, and ability to be able to make high-stakes pass/fail decisions on the basis of a global assessment.

Sources of Error in Performance-Based Assessment

Performance assessments (e.g., OSCE) are prone to a wide variety of measurement errors (Boulet et al., 2003). Case design, SP portrayal, and examiner training can influence the reliability, and subsequently the validity, of a performance-based assessments (Boulet et al., 2003; Downing & Haladyn, 2004; van der Vleuten, 1996; van der Vleuten & Swanson, 1990).

Case Design

Careful and meticulous case design (both case and individual tasks within the case) is one the most efficient strategies for reducing measurement error in performance-based examinations (Boulet et al., 2003; Downing & Haladyn, 2004). The reliability of the assessment will be influenced if the case measures the skills inadequately or the items comprising an assessment checklist are too specific for the content of the case (e.g., evaluating for nystagmus [rapid, involuntary eye movements] in a head injury case) (Boulet et al., 2003).

Downing (2003b) specified a number of principles that should be adhered to when developing performance-based examinations. First, the cases selected must be representative of the course/rotation blueprint (e.g., intensive care). Second, evidence must be presented to demonstrate that faculty content experts have agreed that the cases selected are representative of the area of medicine to be examined (e.g., an Intensive Care Medicine rotation). Evidence must also be provided to display that faculty content experts have designed, reviewed and revised the cases selected, while other faculty content experts have reviewed and critically evaluated the cases. When SPs are involved, all the critical clinical information to be portrayed must be specified, detailed SP training guidelines must be established by faculty content experts, and SP trainers must stringently coach the SPs based on these guidelines. Finally, cases should be developed that are of medium difficulty, as cases that are too easy or too hard will provide very little information on student achievement (Downing, 2004; Downing & Haladyn, 2004).

Case evaluation. One strategy for evaluating whether checklist items are being addressed, or not, is to perform an item analysis. It can be very revealing to know the proportion of candidates receiving or not receiving credit for a checklist item. Determining the number of examinees who do not receive credit for an item can provide information regarding three elements; the SP portrayal might not be adequate, there may be problems with the interpretation of the scoring criteria, or the item may not be appropriate for the skill level of the examinee or the item may be out of the scope of practice for the candidate. Investigation (e.g., videotape review) will be required to identify the specific cause of the discrepancy (Boulet et al., 2003).

Standardized Patients

SP-based examinations are subject to a wide variety of potential measurement errors (Boulet et al., 2003). Measurement error can be introduced via the inconsistency of SP performance (Boulet et al., 2003; van der Vleuten & Swanson, 1990), inaccurate performance by the SP, the choice of SP to portray the case, or the SP's portrayal of the case (Boulet et al., 2003). Measurement errors can also be the result of physical findings that are not related to the case to be portrayed (Boulet et al., 2003; Peitzman, 2001; Williams, 2004). A recent study of potential SPs discovered that over half of the applicants had at least one easy to detect physical abnormality (e.g., heart murmur) and often the applicants were unaware of the abnormality. The concern is that these non-case physical findings could affect performance and diagnostic thinking (Peitzman, 2001). These findings reveal the importance of having physicians examine SPs during the selection process and especially when two or more SPs are being trained for the same case (Peitzman, 2001; Williams, 2004).

The accuracy of SP portrayal is critical during performance assessments when the candidate is gathering information to formulate a diagnosis (or differential diagnoses) and creating the subsequent management plan (Tamblyn et al., 1991). For this reason, careful training of the SP(s) is required to maximize the reliability (reproducibility) of each OSCE case (Adamo, 2003; Boulet et al., 2003; Carraccio & Englander, 2000; Downing & Haladyn, 2004; van der Vleuten & Swanson, 1990). Strube (2000) referred to this as 'standardization'; an attempt to control the measurement conditions so that extraneous sources of error can not influence the scores. Williams (2004) suggested that while refinements in SP training methods had likely led to improvements in the accuracy of SP portrayal, he was unaware of any studies subsequent to Tamblyn (1989) that has evaluated SP accuracy and its potential impact on student performance outcomes.

Exacting specifications must outline all the relevant clinical information to be portrayed by the SPs. Evidence must be provided that the cases have been completely edited and that detailed SP training guidelines and criteria have been created, reviewed by faculty experts and implemented by experienced SP trainers; as all provide evidence of content validity (Downing, 2003b). Reproducibility of the SPs portrayal is fundamental to the claim that the tests are standardized; as in the same case is presented to every examinee (Petrusa, 2002). Properly trained SPs will not vary in their presentation from examinee to examinee (Barrows, 1987). SPs that are not trained to consistently portray the patient in a standardized manner will likely result in different examinees effectively encountering different patients and slightly different patient problems (Downing, 2003b). As a result, there must be documentation that the SP portrayals were monitored closely to ensure all the students experience nearly the same case. Statistics should be presented to demonstrate that a different SP trained on the same case rates examinee performance in the same manner (Downing, 2003b).

Recruitment, thorough training and the development of quality assurance protocols are essential for the continued accurate performance of SPs for clinical evaluation (Adamo, 2003; Glassman et al., 2000). Adamo (2003) stated that while standards and models for standardized patient training have not been documented standards and models do exist in practice. These standards and models can vary from institution to institution and there are often few resources available to conduct quality assurance reviews.

Examiners

Downing (2004) stated that examiner inconsistency or low inter-rater reproducibility is the largest threat to the reliability of an assessment. Standardized training of examiners to use the assessment instruments, regardless of the format(s) (objective criteria checklists and/or global checklists) is critical (Boulet et al., 2003; van der Vleuten & Swanson, 1990; Wass et al., 2001) and extensive examiner training is required if the evaluation is used for summative assessment (Mavis & Henry, 2002).

Trained, experienced physicians are typically considered the most qualified raters of candidate performance on high-stakes examinations (Humphrey-Murto et al., 2005; Martin et al., 1996). Issues such as cost (Boulet et al., 2003; Martin et al., 1996; McLaughlin et al., 2006; van der Vleuten & Swanson, 1990) and availability (recruitment) of physician raters (Boulet et al., 2003; Humphrey-Murto et al., 2005; McLaughlin et al., 2006; van der Vleuten & Swanson, 1990) must be considered when making a decision as to whether a trained non-physician rater can be used to reliably and validly evaluate candidate performance. van der Vleuten and Swanson (1990) suggested when cost and physician availability were issues that well-trained non-physician rater (such as an SP) could be a acceptable alternative.

SP-raters are used to assess candidates challenging the Educational Commission for Foreign Medical Graduates (ECFMG) Clinical Skills Assessment (Boulet, McKinley, Norcini, & Whelan, 2002) as the logistics and cost for processing 5,000 to 10,000 candidates per year is too prohibitive to utilize physician examiners (Whelan, Boulet, McKinley, Norcini, van Zanten, Hambleton, et al., 2005). While the ECFMG has acknowledged that some candidates might not consider an SP evaluation to be an acceptable form of assessment, this objection is countered with the knowledge that physicians are heavily involved with the construction of the case scenarios and the development of the assessment checklists (Boulet et al., 2003).

Generalizability Theory Analysis

Generalizability Theory is an extension of Classical Test Theory (Downing, 2003b) which can be used to identify and estimate multiple sources error using a single analysis (Boulet et al., 2003; Downing, 2004; Shavelson & Webb, 1991; Shavelson, Webb, & Rowley, 1992). Generalizability Theory is a statistical theory that can provide the researcher with a summary coefficient that reflects the dependability of the measurements taken (Shavelson et al., 1991). This co-efficient (Ep^2) is comparable to the reliability coefficient (e.g., Pearson *r*) in classical test theory (Shavelson et al., 1991; Thomas & Nelson, 2001).

A Review of Classical Test Theory

In Classical Test Theory, the observed score is separated into two parts; the true score and the random error of measurement (Crossley, Davies, Humphris, & Jolly, 2002; Downing, 2004; Shavelson & Webb, 1991; Shea & Fortna, 2002; Strube, 2000). Reliability is defined as the ratio of the true score variance to the total score variance (true score plus error) (Crossley, Davies et al., 2002; Dimitrov, 2002; Downing, 2004; Shea & Fortna, 2002).

Downing (2003a) stated that the true score is the most important component of Classical Test Theory and is defined as the "long-run average or mean score" (Downing, 2003a, p. 740). The long run average is the mean of all the scores obtained on a test if the same or an equivalent version of that test was repeated an infinite number of times. In reality, the true score can not be determined so it must be estimated (Downing, 2003a; Shea & Fortna, 2002). The error component of the observed score is assumed to be independent of the true score and is attributable to random noise in the measurement and this random error of measurement can come from multiple sources (e.g. cases, raters and standardized patients) (Shavelson & Webb, 1991; Shavelson et al., 1992).

Three limitations to Classical Test Theory include: 1) the consideration of error as a single unit (Hambleton, 1989; Strube, 2000), 2) whether the examinee's score rises or falls with changes in test difficulty (different testing forms) (Downing, 2003a; Hambleton, 1989), and 3) the limited usefulness of the reliability coefficient if the examinee does not closely represent the population for which the test is intended (Hambleton, 1989).

Introduction to Generalizability Theory

Generalizability Theory (GT) focuses on the dependability of measurements (Shavelson & Webb, 1991; Shavelson & Webb, 2006; Shea & Fortna, 2002) in that a measure's usefulness depends on the degree to which a researcher can generalize that measure into some larger defined universe (Shavelson & Webb, 1991; Shavelson & Webb, 2006). GT is derived from three amendments to Classical Test Theory.

In the first amendment, Shavelson and Webb (2006) explained that GT expanded from the realization that the undifferentiated error in Classical Test Theory was too general a reflection of the actual or potential sources of error in a measurement. For example, measurements in performance-based assessments may contain multiple sources of error (e.g., raters, cases, SPs). These potential sources of error are called facets and the levels of the facets are called conditions (Shavelson & Webb, 2006; Shavelson, Webb, & Rowley, 1989). In GT, multiple sources of error can be estimated separately in one analysis (Shavelson & Webb, 1991; Shavelson & Webb, 2006, Swanson, 1987).

In the second amendment, the conditions of observations on GT are not necessarily parallel. For example, a measurement (e.g., an OSCE case score) is a sample from a universe of all admissible observations (Shavelson & Webb, 1991; Shavelson & Webb, 2006). The candidate's score (the observed behaviour) is defined as the expected value of all the candidate's observed scores in the universe of admissible scores over the long run. This is analogous to the candidate's true score in Classical Test Theory (Shavelson & Webb, 1991; Shavelson & Webb, 2006). Finally, GT can distinguish between relative (or rank order) decisions (norm-referenced) and absolute decision (criterion-referenced) decisions, whereas, CT only focuses on the relative decision.

Facets and Designs

One-facet design. A one-facet design is defined by one potential source of error (Shavelson & Webb, 1991). In an OSCE, for example, the participants' observed scores can be influenced by the cases in the examination. If all the participants (p) manage each case (c) on a multi-case OSCE, it is called a one-facet crossed design and is designated as p x c (participants crossed with cases). A one-facet design has four sources of variability; 1) the differences between the participants (p), 2) the differences in difficulty of the cases (c), 3) the participant by case effect (pc), and 4) random error, or, the residual (Shavelson et al., 1992). The variability as a result of the interaction effect (pc) and the random error can not be separated and are combined together and labelled as the residual (pc, residual) (Burns, 1998; Shavelson & Webb, 1991). A Venn diagram of a single facet crossed design is presented in Figure 2.

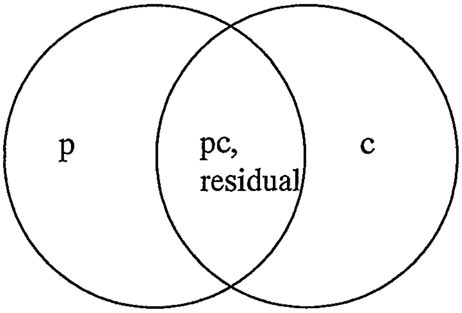

Figure 2. Sources of variability for a single facet, fully crossed design. The main effects are Participants (p) and Cases (c). The interaction effect includes participants crossed with cases, and the residual (pc, residual) (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991).

Two-facet design. Typically, behavioural sciences measurements are complex and have the potential for many sources of error that can influence the observed scores. For this reason, the universe of admissible observations will include more facets. A two-facet design is defined by two potential sources of error, for example, OSCE cases (c) and raters (r) (Shavelson & Webb, 1991). If, in the previous example, all of the participants (p) managed each case (c) and each performance was rated by two raters (r); it is called a two-facet crossed design and is designated as p x c x r (participants crossed with cases crossed with raters). A two-facet crossed design has seven sources of variability; 1) the differences between the participants (p), 2) the differences in difficulty of the cases (c), 3) the differences between the raters (r), 4) the participant by case effect (pc), 5) the participant by rater effect (pr), (6) the rater by case effect (rc), and 7) random error or the residual (Shavelson & Webb, 1991). As with the variability between in interaction effect in a one-fact design, the variability as a result of the interaction effect (pcr) and the random error can not be separated and are combined together as labelled as the residual (pcr, residual) (Burns, 1998; Shavelson & Webb, 1991). A Venn diagram of a two-facet crossed design is presented in Figure 3.

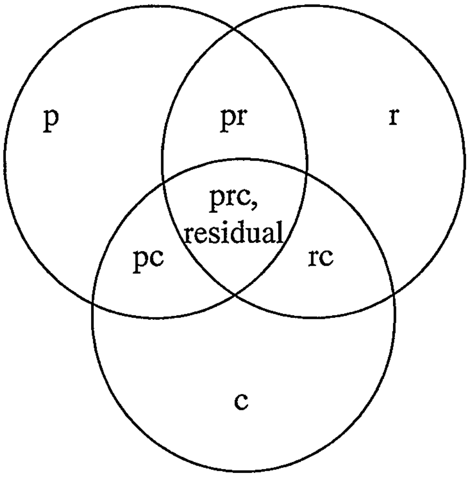

Figure 3. Sources of variability for a two-facet, fully crossed design. The main effects are Participants (p), Cases (c), and Raters (r). The interaction effects include: participants crossed with cases (pc), participants crossed with raters (pr), raters crossed with cases (rc), participants crossed with cases crossed with raters, and the residual (pcr, residual) (Cronbach et al., 1972; Shavelson & Webb, 1991).

Nested design. The feasibility of running a multi-case OSCE with several participants and only having two raters score each and every performance (for example, n = 39 participants, n = 14 cases, equalling n = 546 encounters) is low. Nested designs are a practical solution to feasibility issues such as 'multiple encounters'. A design is called 'nested' if the design meets two conditions; 1) that there are multiple levels of one facet (e.g., Facet A, raters) associated with each level of another facet (e.g., Facet B, cases) and, 2) different levels of the one facet (e.g., Facet A, raters) is associated with each level of the other facet (e.g., Facet B, cases) (Shavelson & Webb, 1991).

Condition 1 is met if two raters (e.g., Facet A, Level 1; Dr. Smith and Dr. Jones) are trained to evaluate all the participants on case #1 (e.g., Facet B, Level 1; chest pain) and two different raters (e.g., Facet A, Level 2; Dr. Black and Dr. White) are trained to evaluate all the participants on case #2 (e.g. Facet B, Level 2; abdominal pain). Condition 2 is met when different levels of Facet A (raters) are associated with each level of Facet B (cases) (Shavelson & Webb, 1991). In other words, each level of each facet is not combined with levels of the other facets (Strube, 2000). This means that Drs. Smith, Jones, Black, and White do not evaluate any other case except the case they have been trained to evaluate.

If all the participants (p) (n = 39) manage each case (c) on a 14-case OSCE (n = 14) and two raters (r) (n = 2) are assigned to each OSCE case (thus, n = 28 raters trained); it is called a two-facet nested design and is designated as [r:c] x p (raters nested into cases [nesting is indicated with a colon] and crossed with participants). A two-facet nested design has five sources of variability; 1) the differences between the participants (p), 2) the differences in difficulty of the cases (c), 3) the participant by case effect (pc), 4) the rater confounded with rater nested into case (c, r:c), and 5) the rater nested into the case (r:c) confounded with the three-way interaction between participant, case, and rater and error or the residual (r:c, pc:r, error) (Cronbach et al., 1972; Shavelson & Webb, 1991). In many assessment designs, the rater (or standardized patient) is nested or confounded with the cases they rate, so it is impossible to directly estimate the error associated with the nested facet (e.g., raters) (Downing, 2004). A Venn diagram of a two-facet nested design is presented in Figure 4.

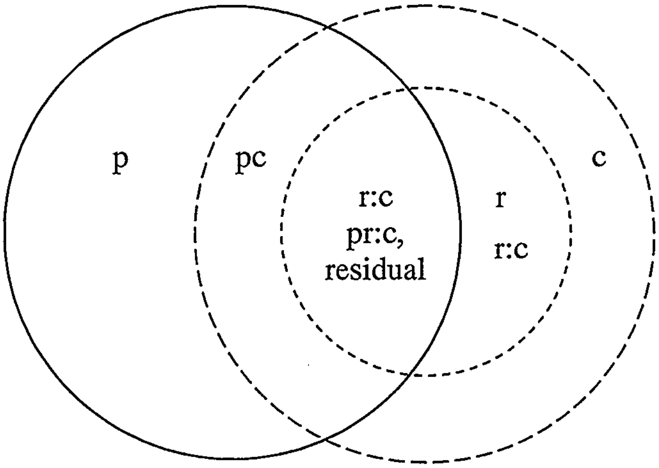

Figure 4. Sources of variability for a two-facet, nested design. The main effects are Participants (p) and Cases (c). The interaction effect is participants crossed with cases (pc). The confounded effects include: raters (r) and raters nested into cases (r and rc), and raters nested into cases (rc) and participants crossed with cases crossed with raters, and error (prc, residual). (Cronbach et al., 1972; Shavelson & Webb, 1991).

Variance Components

Variance components are the "building blocks of generalizability theory" (Brennan, 1994, p. 176). The most common method for calculating the variance components is called the Analysis of Variance (ANOVA) procedure and it involves a series of calculations with the mean squares from an ANOVA table (Brennan, 1994). Cronbach et al. (1972) recommended reporting the percentage of variance accounted for the main facets and their interactions.

What variance components indicate. Variance in the participant facet indicates that there is a difference in skill level amongst the participants, while variance in the case facet indicates that the cases are not of equal difficulty (Boulet et al., 2003). Variance in the interaction effect between participants and cases indicates that there is a difference in how participants managed the cases. For example, a participant might have found some of the cases less difficult to manage (e.g., clinical skills: an evaluation of abdominal pain) yet discovered other cases that were more difficult to manage (e.g., counselling: breaking bad news). It should also be noted that the variability as a result of the interaction effect (pc) and the random error can not be separated and are combined together as labelled as the residual (Shavelson & Webb, 1991). In a nested design with SPs nested into cases (sp:c), a non-zero variance component for sp:c indicates that there are differences in the candidates' scores due to the SPs selected to portray the case (Boulet et al., 2003). Finally, a large residual (error) variance component indicates that there are not enough cases in the OSCE to evaluate the different constructs (e.g., physical assessment skills) (Boulet et al., 2003).

Summary of the Review of Literature

Performance-based assessments, such as the OSCE, are designed to be a controlled assessment of clinical competence by removing the variability introduced by the patient and the examiner. The OSCE fits within the 'Shows how' category of Miller's Pyramid, which reflects the ability of the participant to demonstrate behaviours within a simulated situation. The use of SPs to portray cases was introduced by Barrows in the middle 1960s. SPs can be trained to present a wide variety of cases for training, formative assessment and summative (or high-stakes) assessment. Meticulous training of SPs is required for the accurate and reproducible presentation of a case. Extensive training of examiners to rate candidate performance is also required, as is the development of robust clinical cases. Reliability refers to the consistency or reproducibility of test scores. Case design, SP portrayal, and examiner training can influence the reliability, and subsequently the validity, of performance-based assessments. Generalizability analyses identify and estimate multiple sources error using a single analysis and provides the researcher with a variance component which reflects the sources of variability of the collected scores.

Research Questions

1. What are the sources of variance in clinical performance evaluation, using an Objective Structured Clinical Examination, as determined by generalizability analyses?
2. Is there a difference between physician and non-physician raters in the assessment of clinical competence?

3. Can standardized patients assess communication skills as effectively as physician raters?
4. Can the characteristics, strengths and limitations of an Objective Structured Clinical Examination be identified in order to create a model for the reliable assessment of competency in medicine?
5. What are the consistent errors of measurement introduced by examiners, standardized patients, and case variability?

CHAPTER THREE

Methods

Two studies were undertaken for this dissertation. The purpose of study one was to evaluate the reliability and identify potential sources of measurement error in a fourteen case OSCE used to evaluate the clinical competency of International Medical Graduates (IMGs). The purpose of study two was to evaluate the inter-rater reliability of physician versus non-physician raters in the assessment of clinical competence.[2]

Permission to conduct both studies was granted by the Office of Medical Bioethics, Faculty of Medicine, University of Calgary (Appendix A). The consent form for the release of all personal information and assessment results from the WAAIP project is located in Appendix B.

Western Alliance for the Assessment of International Physicians (WAAIP) Project

The WAAIP project was established to develop a psychometrically sound assessment protocol for IMGs. The mandate was to accelerate the advancement of IMGs whose clinical knowledge and skills were sufficiently competent that it was deemed a residency program would not be required in order to move them into the physician workforce (Western Alliance for Assessment of International Physicians, 2006).

Evaluation was undertaken in a two-step process. In Step A, 39 candidates wrote a 150-item MCQ examination to assess their medical knowledge and participated in a fourteen case OSCE to evaluate their clinical skills. In step (Step B), the top ranked

[2] The present study is focused on the psychometric evaluation of a fourteen case OSCE. The author was not involved in the development and administration of the examination (selection of the candidates, case design, or training of the SPs and physician raters).

25 candidates from Step A were selected for a supervised clinical practice of 3 months duration at sponsoring health care facility in the qualifying candidate's respective province (Violato & Baig, 2006).

The OSCE

Based on a review of current published literature, focus groups, expert input, and best psychometric practices, a table of specifications was developed. The competencies identified on the table of specifications to be assessed included medical expert, professional, collaborator, manager, health advocate, scholar, and communicator. These competencies are in concordance with those advocated by the Accreditation Council for Graduate Medical Education, American Board of Medical Specialties, and the Royal College of Physicians and Surgeons of Canada (Violato & Baig, 2006).

The cases and respective assessment checklists thought to represent this table of specification were derived from a roster of cases previously created by either the Alberta International Medical Graduates (AIMG) program or the British Columbia or Manitoba equivalents. All the cases had been utilized in previous IMG evaluation settings (Violato & Baig, 2006).

A minimum performance level (MPL) for each case was determined utilizing a two-step process called the Ebel procedure. In the first step, a case reviewer (judge) categorizes each checklist item (within the context of the case) on its difficulty (easy, moderate, or hard) and its relevancy (essential, important, or marginal) (Ebel & Frisbe, 1986). In the severe headache case, the item "the physician asks the patient to rate the severity of pain on a scale from 1 to 10" might be categorized as "easy" (difficulty) and "essential" (relevance).

This “difficulty and relevance” protocol creates a 3 x 3 matrix (thus, nine cells), which lays the foundation for the second step of the Ebel procedure. Every cell is assigned a weight (e.g., the “easy” and “essential” cell is weighted at 0.90). Each checklist item assigned to a cell (based on the case reviewer’s categorization) is multiplied by that weight. The weightings are summed and the total is designated the MPL (or cut score) for the case (Violato, Marini, & Lee, 2003). This process was undertaken for each OSCE case (Violato & Baig, 2006).

The Candidates

Candidates must have graduated from a medical school included in the World Health Organization’s directory of medical institutions and had his or her medical degree verified by the Educational Commission for Foreign Medical Graduates International Credentials Services. Each candidate must have met the minimum required standards on the Test of English as Foreign Language (TOEFL) and have passed the Medical Council of Canada Evaluating Examination (MCCEE) (Violato & Baig, 2006). From an original list of one hundred and sixteen candidates, 39 were selected to participate.

Study One

Candidates

Thirty-nine IMGs from British Columbia, Alberta, Saskatchewan, Manitoba, and the Yukon participated in a fourteen case OSCE administered at the Medical Skills Centre, Faculty of Medicine at the University of Calgary.

Assessment of Clinical Competence

The WAAIP project utilized a fourteen case OSCE for evaluating the clinical competence of the IMGs. The major skills to be evaluated included history taking, physical assessment, information sharing, and counselling. A list of the case name, presenting condition, purpose(s) of the case (e.g., history taking), and differential diagnosis and/or treatment (e.g., infection and antibiotics) is located in Appendix C. To maintain the confidentiality of the cases for future use, the name of each case has been changed. For the purposes of this dissertation, each case will be named based on the NATO (North Atlantic Treaty Organization) phonetic alphabet (Alpha, Bravo, Charlie, Delta, Echo, Foxtrot, Golf, Hotel, India, Juliet, Kilo, Lima, Mike, and November).

Procedures

The candidates were randomly assigned to a specific OSCE case as a start point. Prior to entering a testing room, the candidate was provided with a one page outline of the patient (e.g., Kathy Bravo is a 21 year old female), presenting case (e.g., complaining of a severe headache), and the case's requirements (e.g., take a history and perform a physical assessment). If critical to the case, a complete set of baseline vital signs were provided. Upon entering the room, the candidate was given ten minutes to interact with the SP. After ten minutes, the physician rater (who was in the room during the candidate/SP interaction) conducted a two minute post-encounter probe (e.g., What is your differential diagnoses?). A public address system was used to co-ordinate candidate rotation between cases and to standardize the duration of interaction and post-encounter probe phases.

SPs. Twenty-eight actors (two assigned to each case) were trained to portray the patients. Having two SPs per case allowed for two examination tracks (red and blue) to be run concurrently during the morning session. Only one examination track (red) was run in the afternoon session. For the afternoon session, all the SPs from that morning's red track participated while the SPs from the blue track were excused. At no point during the OSCE did an SP trained for one case (e.g., Alpha) present a different case (e.g., Bravo). The Medical Skills Centre Standardized Patient Program, in the Faculty of Medicine, at the University of Calgary trained the actors. [3] A flow chart of SP assignment is located in Figure 5.

Physician Raters. Thirty-one physician raters were used during the course of the OSCE. Twenty-eight raters took part in the morning session (one assigned per case across the two examination tracks). Eleven of the raters from the morning sessions stayed for the afternoon session. Ten of these eleven physician.raters were assigned to a different case from the morning session. Three other physicians joined the morning physicians to complete the afternoon roster of physician raters. [4] A flow chart of physician rater assignment is located in Figure 6.

Assessment Instruments. Each case was evaluated using a case-specific checklist, a five point global rating of overall performance, and a communication skills checklist. The presiding physician rater completed all three of the assessment instruments, while the standardized patient only completed the communication skills checklist.

[3] No information regarding who selected the actors, how long the actors have been in this (or any previous) SP program(s), the duration of training for each case, how many trainers were involved, and whether a physician(s) was involved in the training or the review of the training was provided to the author.

[4] No information regarding how the physician raters were selected, their previous experience in assessing OSCEs and the duration of training for each case (or cases for those raters participating in both the morning and afternoon sessions) was provided to the author.

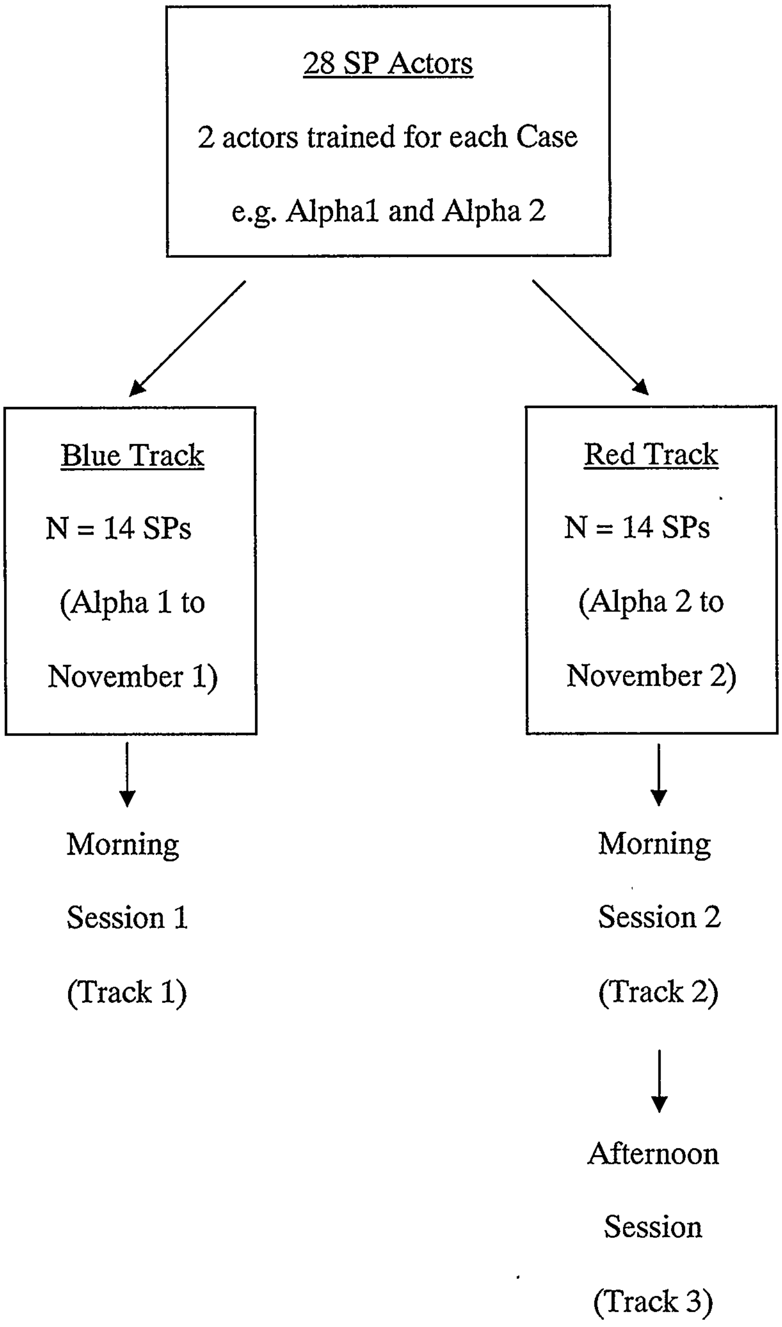

Figure 5. Assignment and schedule for SPs.

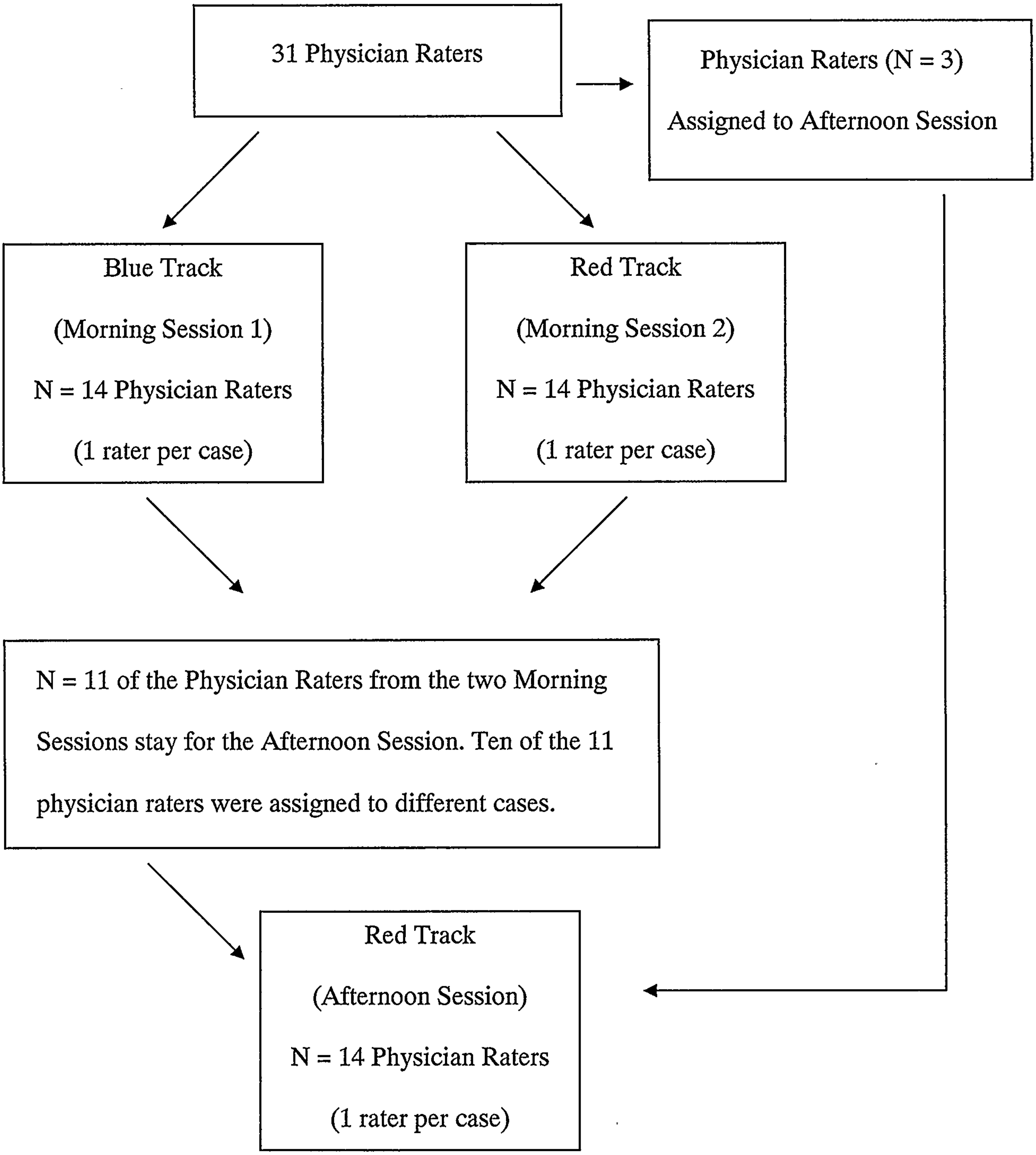

Figure 6. Assignment and schedule for physician raters.

The number of checklist items is case dependent and ranged from 13 items (Delta) to 43 items (Alpha). All of the checklist items were graded 'yes' and 'no' by the physician rater (e.g., "asks when the pain started). Each candidate's checklist score was calculated by adding up the number of 'yes' items addressed.

A global rating scale is located at the bottom of each checklist. The presiding physician rater was instructed to rate the overall performance of the candidate on a one to five scale (1 = Poor; 2 = Borderline Fail; 3 = Borderline Pass; 4 = Good; 5 = Excellent).

A generic 13-item communication checklist was used by both the physician rater and SP. Each item was worded for the perspective of the rater (physician: "the doctor explained the treatment plan to the patient"; SP: "the doctor explained the treatment plan to me"). Each item was rated on a 1 to 5 scale (Strongly Disagree, Disagree, Not Sure, Agree, Strongly agree).

Analysis

Demographics. Demographic analyses of the participants include the frequency distribution and percentages for the following categories: gender (male or female), region of origin (e.g., Asia), whether an internship program was completed (yes or no), and whether a residency program was completed (yes or no). Further demographic analysis of the participants includes the range, mean, and standard deviation for the following categories: age, and number of years since completing medical school. Finally, a demographic analysis for the TOEFL scores will be presented.

The minimum TOEFL eligibility requirements as outlined by the Canadian Residency Matching Service (CaRMS) for IMGs applying for a Canadian residency program are utilized for reference. The minimum internet-based score range from 90

(Alberta) to 100 (Nova Scotia/New Brunswick, Saskatchewan, and British Columbia). The minimum paper-based scores range from 575 (Manitoba) to 600 (Nova Scotia/New Brunswick, Saskatchewan, Alberta and British Columbia). Finally, the minimum computer-based score range from 237 (Newfoundland and Ontario) to 250 (Nova Scotia/New Brunswick, Manitoba, Saskatchewan, Alberta and British Columbia) (Canadian Resident Matching Service, 2008).

The analysis of the TOEFL scores includes a comparison between the mean, standard deviation, minimum and maximum scores for the candidates promoted and not promoted to the three-month clinical rotation and for the candidates who passed and failed the three-month clinical rotation. An ANOVA between groups (promoted versus not promoted; pass versus fail) was calculated to establish whether there is a statistically significant difference between these groups (based on the TOEFL scores).

Descriptive Statistics. A frequency distribution for the number of cases passed, the fail rate for each OSCE, and an evaluation to determine whether a particular assessment track consistently scored higher than the other two tracks across the fourteen cases are presented.

Three scores (checklist, global, and total) were calculated for each case. The total score for a case is the sum of the checklist and global scores. Descriptive statistics include the mean, standard deviation, and minimum and maximum score for the global score (standardized on a 1 to 5 rating) and the checklist and total scores. As the number of checklist items are case-specific, both the checklist and total scores have been transposed into a percent score to allow for a quick comparison between cases (mean overall checklist score [e.g., Alpha = 26.1] divided by the maximum checklist score [e.g., Alpha = 43] multiplied by 100 [60.7%]).

Analysis of Variance (ANOVA). An ANOVA was calculated to determine whether there is a significant difference between the three assessment tracks on the checklist, global, and total scores. A post hoc Scheffe was used to identify where statistically significant differences were located between tracks. An ANOVA was also computed to ascertain whether there is a significant difference in the checklist, global, and total scores between the candidates interacting with one SP team (Blue) versus the other SP team (Red).

Reliability. Internal consistency (Cronbach's alpha) was calculated to evaluate whether items on an case measured the construct they were designed to assess (e.g., communication skills, history taking, or physical assessment).

Internal consistency was calculated with and without the case checklist's post-encounter probe. Each case-specific checklist (with the exception of the November case) is composed of the candidate/SP interaction items (e.g., asks about severity of the pain) and the post-encounter probe (PEP) items (e.g., what is your differential diagnoses?). The number of PEP items varies in number from case to case and are case specific (e.g., not all cases ask for a differential diagnosis). Cronbach's alpha was calculated without the PEP items in order to specifically evaluate the internal consistency of the candidate/SP interaction component of the checklist.

Item Analysis. Boulet et al. (2003) recommended the use of item or case analysis for the statistical evaluation of performance-based examinations (e.g., OSCE). The purpose is to calculate the proportion of candidates who received credit for addressing a specific checklist item (e.g., Bravo case, patient presenting with headache; physical assessment item, check that the pupils are equal and reactive to light). Item analysis is an effective strategy for identifying poorly functioning cases, spotting standardized patients that require further training, or evaluating whether a case that has been designed to assess a specific competency (e.g., physical assessment skills) correlates with other cases that are designed to measure that same competency (Boulet et al., 2003).

While it has been recommended that researchers include the checklist(s) so that opinions and conclusions about the appropriateness of the checklist can be established (Gorter, Rethans, Scherpbier, van der Heijde, Houben, van der Vleuten et al., 2000), the checklists (and the accompanying item analysis) can not be presented in order to protect the confidentiality of the cases for future use.

To provide a general overview of the item analysis, a distribution of items for each case was calculated using a quartile range (0 to 24.9%, 25 to 49.9%, 50 to 74.9%, and 75 to 100%). For example in the Bravo case, Kernig's sign was addressed by 29 of 39 candidates (74.4%), which would place this item in the 50 to 74.9% quartile range. Conversely, the "Social History: Lives alone, has a steady boyfriend" item was addressed by 8 of 39 candidates (20.5%), placing this item in the 0 to 24.9% quartile range.

A second analysis was undertaken to establish whether a candidate could score a 'poor' or 'borderline fail' on the global rating and still pass the case or score a 'borderline pass' on the global rating and fail the case. A third analysis was performed to evaluate to determine whether a candidate could fail a case despite an accurate diagnosis of the clinical complaint or pass a case despite an incorrect diagnosis. Only those cases with one diagnosis were reviewed: Alpha (fever), Bravo (headache), Juliet (hand problem), and Mike (urinary tract).

Correlation (Pearson's r) between each case score and the mean checklist score. Boulet et al. (2003) recommended Pearson's correlation between individual case scores and overall mean scores be calculated to evaluate how well each case is working. They recommended two strategies. The first strategy correlates the case scores (e.g., Alpha) with each candidate's mean score across all the cases (e.g., Alpha through November). The second strategy correlates the case scores (e.g., Alpha) with each candidate's mean score across the remaining cases (e.g., the mean score for Bravo through to November, thus excluding Alpha). A negative value indicates that low ability candidates scored high on the case or high ability candidates scored low on the case. A value close to zero is also considered unacceptable. An ANOVA was then calculated to establish whether there is a significant difference between the coefficient calculated when both evaluation strategies are compared.

Generalizability Analyses. Generalizability Theory is an extension of Classical Test Theory (Downing, 2003b) that can be used to identify and estimate multiple sources error using a single analysis (Boulet et al., 2003; Downing, 2004; Shavelson & Webb, 1991). Generalizability analyses provides the researcher with a variance component for each facet (the percent variance is then calculated) and a summary coefficient, known as Ep^2, that is comparable to the reliability coefficient Pearson *r* in classical test theory (Rogers et al., 2000; Shavelson et al., 1992).

A series of Generalizability analyses were calculated for the checklist, global and total scores using both a one-facet crossed and two-facet nested and assigned designs.

G_String_II (G_String version 3.1.1), a Windows program based on Robert L. Brennan's urGENOVA, was used to calculate the variance component for each facet and the Generalizability co-efficient (Ep^2) for both the crossed and nested designs.

In order to calculate the variance components for the assigned designs (physician raters and SP/physician rater team), SPSS (Version 14.0.1) was utilized to calculate the variance component for each facet. The Ep^2 for the assigned designs was calculated using the formula derived from the G_String_II program. The critical value of the reliability coefficient is set at .90 (high-stakes assessment) (Downing, 2004; Shea & Fortna, 2002).

The percent variance was calculated for each analysis (e.g., global score; SPs nested into case). In this example, the four variance components (participants, cases, participants x case, and SPs nested into case) were summed. Each variance component was then divided by the summed value and subsequently multiplied by 100 to get the percent variance component.

One-facet analyses (participants x case) was calculated for the checklist, global, and total scores. Two-facet nested analyses were performed to evaluate whether there were differences in the candidates' scores due to the SPs selected to portray the case. Two SPs were trained for each case and at no point did an SP trained for one case present a different case. This meets the criteria outlined in Shavelson and Webb (1991) for 'nesting'. Analyses for all the assessment formats (checklist, global, and total score) for the three examination tracks (two morning and one afternoon) and between the SPs (one team portrayed the case in the morning only, while the other team portrayed the case in the morning and afternoon) were undertaken.

Two-facet assigned analyses were performed to establish whether there were differences in the candidates' scores due to the physician raters assigned to a case. Thirty-one physician raters were trained for the OSCE. The physician raters were not nested within a specific case. Ten of the physician raters evaluated one case in the morning session and a different case in the afternoon session, while one physician rater evaluated the same case in both the morning and the afternoon session. Three new physician raters were assigned to the afternoon session.

Two-facet assigned analyses were performed to establish whether there were differences in the candidates' scores due to the Rater/SP combinations assigned to a case. Thirty-one physician raters and twenty-eight SPs participated, which resulted in forty-two Rater/SP combinations.

Communication Skills. The physician rater and the SP rated the communication skills of the candidates using a 13-item checklist. The items were identical, however, they were written for the different perspectives of the raters (e.g., Physician rater: The doctor wanted to understand how the patient saw things; SP: The doctor wanted to understand how I saw things). Each item was scored on a one to five rating (1 = strongly disagree, 2 = disagree, 3 = not sure, 4 = agree, and 5 = strongly agree). The ratings were reversed for questions 3 and 4 (the doctor took no notice of what the patient felt and the doctor's response was fixed and automatic).). A two-facet nested design was utilized to calculate the variance components. The analyses were performed using SPSS (Version 14.0.1).

Internal consistency measures (Cronbach's alpha) for the thirteen-item communication checklist were evaluated for both the physician raters and SPs scores. A generalizability analysis was performed on each checklist item to calculate the variance components for participants, cases, participants crossed with cases, SPs nested into case, and physician raters assigned to case.

An series of ANOVAs were calculated to identify where differences between scores (e.g., SPs or assessment tracks) were located. The limitation of generalizability analysis is that the percent variance components calculated for the communication data only provides evidence that the SPs rating of the candidates' communication skills did influence the scores but generalizability analysis can not identify where the differences are located. Boulet et al. (2003) recommended that a comparison between the mean scores is one strategy to identify where differences might be located.

Study Two

During the course of the WAAIP OSCE, all of the candidates were videotaped using an unobtrusive, built-in audio and video recording system. The objective was to evaluate the inter-rater reliability between the presiding physician and SP raters during the WAAIP examination and the three videotape raters (a physician and two non-physicians).

Candidates

For the purposes of study two, fifteen of the original WAAIP candidates were randomly selected and their videotaped performance on each OSCE case was evaluated by one physician and two non-physician raters.

Participants (Raters)

One physician and two non-physicians were recruited for study two. The physician was a Ph.D. Candidate in the Department of Medical Sciences (Medical

Education), Faculty of Medicine, at the University of Calgary. Non-physician rater one was a Ph.D. Candidate in the Faculty of Kinesiology at the University of Calgary. Non-physician rater two (NP2) was a Masters Candidate in the Faculty of Education at the University of Calgary.

Procedures

Study two took place between February and April, 2006 at the Medical Skills Centre, Faculty of Medicine, University of Calgary. The physician rater for the videotape assessment was responsible for training the two non-physician raters on the necessary evaluation protocols for a case (e.g., the appropriate procedures for ausculating the heart) and how to use that case's assessment instrument (checklist and global rating scale).

The three videotape raters practiced using the assessment instruments by reviewing videotapes of the candidates not selected for study two. When the non-physician raters stated they were comfortable with using the assessment instruments, the process of evaluating the fifteen selected candidates on that case commenced. The duration of time required to train the non-physician candidates varied depending on the complexity of the case (e.g., history and counselling versus history and physician assessment) and the number of checklist items (e.g., Delta [13 items] versus Alpha [43 items]). Training duration typically ranged from 30 to 60 minutes for each case. A candidate's performance was assessed without pausing or reviewing the videotape and without discussion between the videotape raters. Upon completing the fifteen assessments for a specific case, training would begin for the next case.

Analysis

Demographics. Demographic analysis of the participants includes the frequency distribution and percentages for the following categories: gender (male or female), region of origin (e.g., Asia, Middle East or South American), whether an internship was completed (yes or no), whether a residency was completed (yes or no). Further demographic analysis of the participants includes the range, mean, and standard deviation for the following categories: age and number of years since completing medical school.

Descriptive Statistics. The physician raters who evaluated the candidates during the course of the WAAIP OSCE will be identified collectively as WP (WAAIP physician). The videotape physician rater will be designated as VP, while the two video non-physician raters will be identified as NP1 and NP2. The mean and standard deviation on each case for the WAAIP physician raters, videotape physician rater, and both non-physician raters are reported.

Analysis of Variance (ANOVA). An ANOVA was calculated to determine whether there is a significant difference between the physician raters and non-physician raters on the three assessment formats (checklist, global, and total scores).

Intra-Class Correlation (ICC). ICC calculations are based on the intra-class correlation procedures presented in Shea and Fortuna (2002) and Streiner and Norman (1989). The calculation involves the use of a repeated measures ANOVA that calculates the mean squares (MS) and subsequently the variance components for candidates, raters, and the residual. The reliability coefficient was calculated using the variance component for the candidates in the numerator, divided by the sum of variance components for candidates, raters, and residual in the denominator. Shea and Fortna (2002) recommended the inclusion of the raters' variance component as the raters are only a sample of all possible raters that could be employed in OSCE testing. The ICC was calculated for the physician and non-physician raters on all three assessment formats.

Generalizability Analyses. Generalizability Theory is an extension of Classical Test Theory (Downing, 2003b) that can be used to identify and estimate multiple sources error using a single analysis (Boulet et al., 2003; Downing, 2004; Shavelson & Webb, 1991). Generalizability analyses were calculated for both the physician and non-physician raters' score on the three assessment formats.

A series of Generalizability analyses were performed, using a two-facet crossed design (Candidates x Cases x Raters) for the physician and non-physician data. The non-physician scores represent a truly crossed design as both non-physicians raters evaluated the fifteen candidates on all fourteen cases. The physician data is not truly a crossed design. While the videotape physician rater evaluated all fifteen candidates on all fourteen cases, the same can not be said for the corresponding WAAIP physician rater data. G_String_II (G_String version 3.1.1), a Windows program based on Robert L. Brennan's urGENOVA, was used to calculate the variance components and the Generalizability co-efficient (Ep^2).

Communication Skills. Internal consistency measures (Cronbach's alpha) for the thirteen-item communication checklist were evaluated for the physician raters, non-physician raters, and SPs scores.

Summary of Research Question Analyses

A summary of the statistical analyses used to evaluate each research question is presented in Table 1.

Table 1

A summary of the statistical analyses used for each research question

Research Question	Statistical Analyses
What are the sources of variance in clinical performance evaluation, using an Objective Structured Clinical Examination, as determined by generalizability analyses?	- SPs: two-facet nested analyses - Physician Raters: two-facet assigned analyses - Rater/SP combinations: two-facet assigned analyses
Is there a difference between physician and non-physician raters in the assessment of clinical competence?	- Analysis of Variance - Two-facet crossed analyses - Intra-class coefficients
Can standardized patients assess communication skills as effectively as physician raters?	- Cronbach's alpha - Generalizability analyses
Can the characteristics, strengths and limitations of an Objective Structured Clinical Examination be identified in order to create a model for the reliable assessment of competency in medicine?	- Generalizability analyses
What are the consistent errors of measurement introduced by examiners, standardized patients, and case variability?	- Generalizability analyses

CHAPTER FOUR

Results

Two studies were undertaken for this dissertation. The purpose of study one was to evaluate the reliability and identify potential sources of measurement error in a fourteen case OSCE used to evaluate the clinical competency of International Medical Graduates (IMGs). The purpose of study two was to evaluate the inter-rater reliability of physician versus non-physician raters in the assessment of clinical competency.

The results for study one will be presented first. The demographics of the IMGs will be followed by a statistical analysis of the OSCE results (descriptive statistics, correlation analysis, analysis of variance, item analysis, and generalizability analyses). The OSCE results will be followed by an analysis of the 13-item communication skills checklist (descriptive statistics, internal consistency, and generalizability analysis).

Section two will focus on the results of the inter-rater study comparing physician versus non-physician raters. Demographics of the randomly selected IMGs will be followed by a statistical analysis of the OSCE results (descriptive statistics, correlation analysis, analysis of variance, internal consistency, intra-class coefficient, and generalizability analyses. The OSCE results will be followed by an analysis of the 13-item communication skills checklist (descriptive statistics and internal consistency).

Study One

Thirty-nine IMGs from British Columbia, Alberta, Saskatchewan, Manitoba, and the Yukon participated. As all thirty-nine participants were vying for one of twenty-five positions in a three month clinical rotation, they will be referred to, for the remainder of the dissertation, as 'the candidate' or 'the candidates'.

Demographics

Test of English as a Foreign Language (TOEFL). Two of the thirty-nine candidates identified English as their first language. The remaining thirty-seven candidates were required to provide their TOEFL scores. Three examination formats are available (in brackets, the number of candidates presenting TOEFL scores in the respective format): internet (N = 1), paper-based (N = 3), and computer-based (N = 33).

The internet-based candidate exceeded all provincial CaRMS requirements. Two of the three paper-based candidates surpassed the minimum CaRMS requirement. At the risk of identifying the internet and paper-based TOEFL format candidates based on promotion to the three-month clinical rotation, only the scores and promotion status for the computer-based TOEFL examination format will be presented.

Thirty of the thirty-three computer-based candidates exceeded the minimum CaRMS requirement of 237, while twenty-three candidates surpassed the minimum CaRMS requirement of 250.The mean TOEFL score for the twenty-one (computer-based TOEFL) candidates promoted to the three-month clinical rotation equalled 256.7 (SD = 18.3 years; Minimum = 220; Maximum = 287). The mean TOEFL score for the 12 (computer-based TOEFL) candidates not promoted to the three-month clinical rotation equalled 251.7 (SD = 13.5 years; Minimum = 223; Maximum = 267). There is no significant difference between the TOEFL scores, based on the promotion to the clinical rotation (*F* [1, 31] = .801, p < .378).

Of the twenty-one candidates (computer-based TOEFL assessment) promoted to the three-month clinical rotation, 13 passed and 8 failed. The mean TOEFL score for the 13 candidates who did pass the three-month clinical rotation equalled 260.5 (SD = 16.2 years; Minimum = 237; Maximum = 287). The mean TOEFL score for the 8 (computer-candidates who did not pass the three-month clinical rotation equalled 250.4 (SD = 20.8 years; Minimum = 220; Maximum = 267). There is no significant difference between TOEFL scores, based on the promotion to the clinical rotation (*F* [1, 19] = 1.578, p < .224).

Age and Gender. Nineteen males (48.7%) and twenty females (51.3%) were evaluated. The mean age of the sample (n = 39) is 41.4 years (SD = 6.6 years). There is no significant difference between age, based on gender (*F* [1, 37] = .177, p < .676). Summary statistics for age and gender are located in Table 2.

Table 2

Descriptive statistics for age (in years) and gender of the candidates (N = 39)

Category	Number	M	SD	Min.	Max.	Range
Male	19	41.8	7.6	29	55	26
Female	20	41.0	5.5	29	52	23
Total	39	41.4	6.5	29	55	26

Country of Origin. Twenty-two countries of origin are represented in the sample. While a few countries were represented by multiple candidates most countries were only represented by one. Due to the risk of identifying a candidate based on country of origin and date of the WAAIP examination, only data for general regions is presented. Fifteen

(38.4%) candidates originated from Asia, six (15.4%) from Eastern Europe, and fifteen (38.4%) from the Middle East. Three candidates (7.8%) have been categorized as 'Other' to avoid potential identification.

Years Since Completing Medical School Training. All candidates possess an MD degree. The mean number of years since completing medical school (N = 39) is 16.2 years. There is no significant difference between gender (*F* [1, 37] = .092, p < .764). The descriptive statistics for the number of years since completing medical school are located in Table 3.

Table 3

Descriptive statistics for the number of years since completing medical school

Category	Number	M	SD	Min.	Max.	Range
Male	19	16.6	7.7	4	30	26
Female	20	15.9	6.3	5	29	24
Total	39	16.2	6.9	4	30	26

Internship and Residency Training. Thirty candidates (76.9%) have completed an Internship program, three have not (7.7%), and information was not provided by six candidates (15.4%). Twenty of the candidates (66.7%) whom have completed an Internship have also completed a Residency program, in their country of origin.

OSCE Results

The WAAIP OSCE was administered in one day using three examination tracks (two in the morning and one in the afternoon). Two SPs were trained for each OSCE case to create two rosters of SPs (Blue and Red). The Blue team portrayed their respective case in one morning session only. The Red team Red presented their respective case in the second morning session and the afternoon. Results will be presented based on both track and SP assignments.

Clinical performance on each case was assessed by a physician rater using a case-specific checklist and a global rating score. Three scores per case will be presented (checklist, global, and total score). Communication skills were evaluated by the physician rater and the SP with a thirteen item checklist. The statistical analyses were performed using the Statistical Package for the Social Sciences (SPSS) Version 14.0.1.

Distribution for the number of cases passed. The total number of OSCE cases passed ranged from two to fourteen cases with the distribution positively skewed. Only one candidate passed all fourteen cases, with nine candidates passing thirteen and eleven passing twelve cases. On the lower end, one candidate passed two cases, one passed four, and one passed seven cases. A frequency distribution based on the total number of cases passed is presented in Figure 7.

Fail rate per OSCE case. From a total of 546 candidate/SP interactions, there were 122 failures (22.3%) with six cases having a failure rate above 25%. The highest fail rate occurred in the Charlie case where 19 of 39 (49%) candidates failed. Two differential diagnoses were identified for this case (due to test confidentiality, they can not be identified). A total of four candidates (10.3%) came up with the first diagnosis, while nine candidates (23%) proposed that the patient could have the second diagnosis.

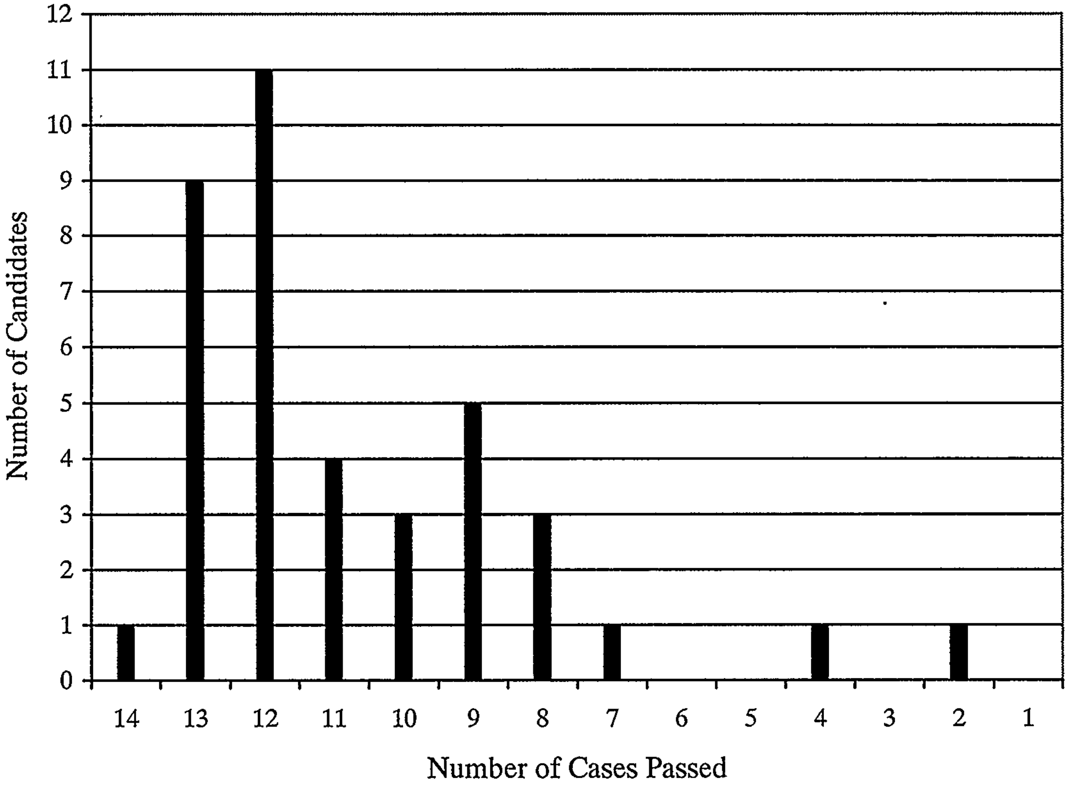

Figure 7. The distribution for number of cases passed.

The number and percentage (in brackets) of failures per case are as follows: Alpha: 9 (23.1%); Bravo: 6 (15.4%); Charlie: 19 (48.7%); Delta: 7 (17.9%); Echo: 10 (25.6%); Foxtrot: 10 (25.6%); Golf: 7 (17.9%); Hotel: 13 (33.3%); India: 8 (20.5%); Juliet: 14 (35.9%); Kilo: 3 (7.7%); Lima: 11 (28.2%); Mike: 7 (17.9%); November: 3 (7.7%).

Track rankings. An investigation was undertaken to determine whether a particular assessment track consistently scored higher than the other two tracks on the fourteen cases. The mean total score for each assessment track on every case was ranked 1^{st}, 2^{nd}, or 3^{rd}. The results indicated that Tracks 1 and 2 were ranked first overall on five cases with Track 3 ranked first overall on four cases. These findings indicate that each track was composed equally of participants with varying levels (e.g., high, medium, low) of clinical competency. The complete track ranking results are presented in Table 4.

Checklist score. The mean score and standard deviation for the checklist score based on tracks and overall is presented in Table 5. Transposing the mean overall checklist score into a percent score reveals that eight cases have a mean score over 60% (standard deviation in brackets): Alpha, 60.7(12.9); Bravo, 62.2(13.7); Delta, 60.2(14.2); Echo, 60.8(9.4); Golf, 61.5(8.8); Kilo, 64.5(7.3); Lima, 60.0(12.3); Mike, 64.2(10.4).

Six cases have a mean score below 60% (mean and standard deviation in brackets): Charlie, 50.2(9.8); Foxtrot, 49.5(15.8); Hotel, 52.8(8.1); India, 56.3(12.7); Juliet, 56.6(10.5); November = 58.1(11.3). The mean percent checklist score for all candidates (n = 39) across the 546 candidate/SP interactions was 58.1(11.3).

A more comprehensive presentation of the checklist scores (including the mean, standard deviation, standard error of measurement, minimum score, maximum score, and range) for the three assessment tracks and overall is located in Appendix D. Descriptive statistics for the checklist scores, based on SP assignment is located in Appendix E.

Global score. There were 546 candidate/SP interactions during the WAAIP OSCE with 57 interactions (10.4%) rated as 'Poor', 80 (14.7%) rated as 'Borderline Fail', 175 (32.1%) rated as 'Borderline Pass', 204 (37.4%) rated as 'Good', and 30 interactions (5.5%) rated as 'Excellent'. The mean and standard deviation for the global score based on the three assessment tracks and overall is presented in Table 6.

A more extensive presentation of the global scores (including the mean, standard deviation, standard error of measurement, minimum score, maximum score, and range) for each assessment tracks and overall is located in Appendix F. Descriptive statistics for

Table 4

The distribution of OSCE cases based on the ranking (1st, 2nd, and 3rd) between the three examination tracks (1, 2, and 3)

Track	Case Ranking		
	1st (mean total score)	2nd (mean total score)	3rd (mean total score)
Blue	Delta (12.92)*	Charlie (15.6)	Alpha (26.3)
(n = 13)	Hotel (25.3)	Kilo (25.7)	Bravo (25.2)
Morning	India (25.1)	Lima (24.2)	Echo (24.92)*
Session 1	Juliet (24.5)	Mike (21.4)	Foxtrot (9.5)
	November (18.4)		Golf (21.9)
Red	Bravo (30.9)	Alpha (29.5)	Delta (8.67)*
(n = 15)	Charlie (18.1)	Golf (24.3)	Hotel (22.9)
Morning	Echo (25.33)*	India (18.9)	Lima (22.2)
Session 2	Foxtrot (11.5)	Juliet (22.3)	Mike (20.9)
	Kilo (26.7)	November (17.3)	
Red	Alpha (32.2)	Bravo (28.8)	Charlie (15.3)
(n = 11)	Golf (24.9)	Delta (12.91)*	India (18.6)
Afternoon	Lima (28.2)	Echo (25.27)*	Juliet (21.7)
Session	Mike (22.7)	Foxtrot (10.6)	Kilo (25.5)
		Hotel (24.3)	November (15.0)

Note: * indicates that the mean had to be extended into the hundredth decimal point to break a tie between tracks.

Table 5

The mean and standard deviation (in brackets) for the checklist score based on Track (1, 2, and 3) and overall

Case	Maximum	Track 1[a]	Track 2[b]	Track 3[b]	Overall
	Score Per Case	(n = 13)	(n = 15)	(n = 11)	(n = 39)
Alpha	43	24.3 (4.2)	25.9 (5.6)	28.6 (6.3)	26.1 (5.6)
Bravo	41	22.8 (5.6)	27.7 (4.4)	25.6 (6.2)	25.5 (5.6)
Charlie	28	13.5 (2.4)	15.3 (2.7)	13.0 (2.7)	14.1 (2.7)
Delta	13	8.7 (1.2)	6.4 (1.9)	8.7 (1.2)	7.8 (1.9)
Echo	36	21.5 (3.7)	21.8 (3.8)	22.5 (2.7)	21.9 (3.4)
Foxtrot	16	7.6 (2.7)	8.3 (2.1)	7.7 (3.0)	7.9 (2.5)
Golf	33	19.0 (2.6)	20.8 (2.4)	21.2 (3.6)	20.3 (2.9)
Hotel	40	21.9 (2.7)	19.7 (3.2)	22.1 (3.5)	21.1 (3.3)
India	32	21.7 (3.7)	16.2 (3.4)	16.2 (2.0)	18.0 (4.1)
Juliet	35	21.3 (4.1)	19.3 (2.9)	18.7 (3.8)	19.8 (3.7)
Kilo	35	22.5 (2.9)	22.8 (2.0)	22.5 (3.0)	22.6 (2.6)
Lima	36	21.7 (3.4)	19.2 (4.0)	24.7 (4.4)	21.6 (4.4)
Mike	28	17.6 (2.8)	17.3 (2.7)	19.4 (3.4)	20.0 (2.9)
November	23	14.8 (2.4)	13.4 (2.1)	11.6 (2.5)	13.4 (2.6)

[a] Track 1 (SP Blue Team)

[b] Tracks 2 and 3 (SP Red Team)

Table 6

The mean and standard deviation (in brackets) for the global score by Track (1, 2, and 3) and overall

Case	Maximum	Track 1[a]	Track 2[b]	Track 3[b]	Overall
	Score per Case	(n = 13)	(n = 15)	(n = 11)	(n = 39)
Alpha	5	2.0 (0.7)	3.7 (0.8)	3.6 (1.4)	3.1 (1.2)
Bravo	5	2.5 (1.1)	3.1 (1.0)	3.2 (1.3)	2.9 (1.1)
Charlie	5	2.1 (1.0)	2.9 (0.7)	2.3 (1.0)	2.4 (1.0)
Delta	5	4.2 (0.4)	2.3 (1.2)	4.2 (0.6)	3.5 (1.3)
Echo	5	3.4 (0.8)	3.5 (0.5)	2.8 (1.1)	3.3 (0.8)
Foxtrot	5	1.9 (1.3)	3.1 (1.2)	2.9 (1.3)	2.7 (1.3)
Golf	5	2.9 (1.0)	3.5 (0.5)	3.7 (1.1)	3.4 (0.9)
Hotel	5	3.4 (0.7)	3.3 (0.8)	2.2 (1.0)	3.0 (1.0)
India	5	3.4 (0.9)	2.7 (1.5)	2.5 (1.0)	2.9 (1.2)
Juliet	5	3.2 (1.0)	3.1 (0.7)	3.0 (1.0)	3.1 (0.9)
Kilo	5	3.2 (0.6)	3.9 (0.3)	3.0 (0.6)	3.4 (0.6)
Lima	5	2.5 (1.1)	3.0 (1.2)	3.5 (0.7)	2.9 (1.1)
Mike	5	3.8 (0.7)	3.7 (0.6)	3.4 (0.8)	3.6 (0.7)
November	5	3.6 (0.8)	3.9 (0.8)	3.4 (1.1)	3.6 (0.9)

[a] Track 1 (SP Blue Team)

[b] Tracks 2 and 3 (SP Red Team)

the global scores, based on SP assignment, is located in Appendix G.

Total score. The mean score and standard deviation for the total score based on tracks and overall is presented in Table 7. Transposing the mean overall total score into a percent score reveals that nine cases have a mean score over 60% (mean and standard deviation in brackets): Alpha, 60.9(13.4); Bravo, 61.8(14.2); Delta, 62.7(16.9); Echo, 61.4(9.6); Golf, 62.3(9.5); Kilo, 65.0(7.4); Lima, 60.0(12.5); Mike, 65.4(10.4); November, 60.7(11.8).

Five cases have a mean score below 60% (standard deviation in brackets): Charlie, 49.9(10.6); Foxtrot, 50.4(16.3); Hotel, 53.6(8.4); India, 56.4(13.4); Juliet, 57.2(11.0). The mean percent total score for all candidates ($n = 39$) across the 546 candidate/SP interactions was 59.1(12.9).

A more comprehensive presentation of the total scores (including the mean, standard deviation, standard error of measurement, minimum score, maximum score, and range), for each track and overall is located in Appendix H. Descriptive statistics for the total scores, based on SP assignment, can be found in Appendix I.

Analysis of Variance (ANOVA).

An ANOVA was calculated to determine whether there are significant differences between the three assessment tracks (on the checklist, global, and total scores) and between the two SP teams (on the checklist, global, and total scores).

ANOVA Checklist score (Three Tracks). There was a significant difference between the three tracks on four of fourteen cases: Delta (F [2, 36] = 10.993, $p < .0001$), India (F [2, 36] = 12.795, $p < .0001$), Lima (F [2, 36] = 6.313, $p < .004$), and November

Table 7

The mean and standard deviation (in brackets) for the total score by Track (1, 2, and 3) and overall

Case	Maximum Score Per Case	Track 1[a]	Track 2[b]	Track 3[b]	Overall
		(n = 13)	(n = 15)	(n = 11)	(n = 39)
Alpha	48	26.3 (4.7)	29.5 (6.1)	32.2 (7.6)	29.2 (6.4)
Bravo	46	25.2 (6.5)	30.9 (5.2)	28.8 (7.2)	28.4 (6.5)
Charlie	33	15.6 (3.2)	18.1 (3.3)	15.3 (3.5)	16.5 (3.5)
Delta	18	12.9 (1.5)	8.7 (3.0)	12.9 (1.7)	11.3 (3.0)
Echo	41	24.9 (4.3)	25.3 (4.2)	25.3 (3.6)	25.2 (3.9)
Foxtrot	21	9.5 (3.7)	11.5 (2.4)	10.6 (4.2)	10.6 (3.4)
Golf	38	21.9 (3.2)	24.3 (2.8)	24.9 (4.6)	23.7 (3.6)
Hotel	45	25.3 (3.2)	22.9 (3.8)	24.3 (4.3)	24.1 (3.8)
India	37	25.1 (4.3)	18.9 (4.6)	18.6 (2.6)	20.9 (5.0)
Juliet	40	24.5 (4.9)	22.3 (3.5)	21.7 (4.7)	22.9 (4.4)
Kilo	40	25.7 (3.4)	26.7 (2.1)	25.5 (3.5)	26.0 (2.9)
Lima	41	24.2 (3.8)	22.2 (5.0)	28.2 (4.9)	24.6 (5.1)
Mike	33	21.4 (3.1)	20.9 (3.2)	22.7 (4.1)	21.6 (3.5)
November	28	18.4 (3.1)	17.3 (2.7)	15.0 (3.5)	17.0 (3.3)

[a] Track 1 (SP Blue Team)

[b] Tracks 2 and 3 (SP Red Team)

(*F* [2, 36] = 29.261, p < .009). A post hoc Scheffe was utilized to determine where the difference(s) are located. The results are presented in Table 8. The ANOVA results for all fourteen cases are located in Appendix J.

ANOVA Checklist score (SP assignment). There was a significant difference between the scores on five of fourteen cases. Bravo (*F* [1, 37] = 5.030, p < .031), Delta (*F* [1, 37] = 4.772, p < .035), Golf (*F* [1, 37] = 4.270, p < .046), India (*F* [1, 37] = 26.301, p < .001), and November (*F* [1, 37] = 6.637, p < .014). The ANOVA results for all fourteen cases are located in Appendix K.

ANOVA Global score (Three Tracks). There was a significant difference between the three tracks on five of fourteen cases: Alpha (*F* [2, 36] = 11.335, p < .0001), Delta (*F* [2, 36] = 23.316, p < .0001), Foxtrot (*F* [2, 36] = 3.596, p < .038), Hotel (*F* [2, 36] = 7.881, p < .001), and Kilo (*F* [2, 36] = 12.366, p < .0001). A post hoc Scheffe was used to determine where the difference(s) are located with the results presented in Table 9. The ANOVA results for all fourteen cases are located in Appendix L.

ANOVA global score (SP assignment). There was a significant difference between the scores on four of fourteen cases. Alpha (*F* [1, 37] = 23.145, p < .001), Delta (*F* [1, 37] = 8.512, p < .006), Foxtrot (*F* [1, 37] = 7.139, p < .011), and Golf (*F* [1, 37] = 5.027, p < .031). The ANOVA results for all fourteen cases are located in Appendix M.

Total score (Three Tracks). There was a significant difference between the three tracks on four of fourteen cases: Delta (*F* [2, 36] = 16.385, p < .0001), India (*F* [2, 36] = 10.544, p < .0001), Lima (*F* [2, 36] = 5.413, p < .009), and November (*F* [2, 36] = 3.684, p < .035). A post hoc Scheffe was run to determine where the difference(s) between

Table 8

The location of significant differences between tracks on the checklist scores

Case Name	Case Problem	Differences Between Tracks	p value
Delta	Consent	Tracks 1 and 2	p < .001
		Tracks 2 and 3	p < .002
India	Fatigue	Tracks 1 and 2	p < .0001
		Tracks 1 and 3	p < .001
Lima	Vomiting	Tracks 2 and 3	p < .004
November	Cardiac Counselling	Tracks 1 and 3	p < .009

Note: Track 1 (Blue Team SPs) and Tracks 2 and 3 (Red Team SPs)

tracks are located. The results are presented in Table 10. The ANOVA results for all fourteen cases, for the total scores, are located in Appendix N.

Total score (SP assignment). There was a significant difference between the scores on five of fourteen cases. Alpha (*F* [1, 37] = 4.306, p < .045), Bravo (*F* [1, 37] = 5.099, p < .030), Delta (*F* [1, 37] = 6.534, p < .015), Golf (*F* [1, 37] = 4.992, p < .032), and India (*F* [1, 37} = 21.641, p < .001). The ANOVA results for all fourteen cases are located in Appendix O.

Internal Consistency Reliability Coefficient (Cronbach's Alpha)

Cronbach's alpha was used to evaluate the internal consistency of each case-specific checklist. Only one case scored higher than 0.70, three cases scored between 0.60 and 0.69, three scored between 0.50 and 0.59, four scored between 0.40 and 0.49,

Table 9

The location of significant differences between tracks based on global scores

Case Name	Case Problem	Differences Between Tracks	p value
Alpha	Fever	Tracks 1 and 2	p < .0001
		Tracks 1 and 3	p < .003
Delta	Consent	Tracks 1 and 2	p < .0001
		Tracks 2 and 3	p < .0001
Foxtrot	Cancer Metastasis	Tracks 1 and 2	p < .048
Hotel	Flu Symptoms	Tracks 1 and 3	p < .004
		Tracks 2 and 3	p < .0001
Kilo	Personality Changes	Tracks 1 and 2	p < .003
		Tracks 2 and 3	p < .0001

Note: Track 1 (Blue Team SPs) and Tracks 2 and 3 (Red Team SPs)

cases ranged between 0.30 and 0.39, with the Cronbach's alpha for the remaining case equalling 0.19.

Each case-specific checklist (with the exception of the November case) is composed of the candidate/SP interaction items and the post-encounter probe items. Cronbach's alpha was calculated with the post-encounter probe items removed in order to specifically evaluate the internal consistency of the candidate/SP interaction components of the checklist. Only one case scored higher than 0.70, two cases scored between 0.60 and 0.69, four scored between 0.50 and 0.59, two scored between 0.40 and 0.49, two

Table 10

The location of significant differences between tracks on the total scores

Case Name	Case Problem	Differences Between Tracks	p value
Delta	Consent	Tracks 1 and 2	p < .0001
		Tracks 2 and 3	p < .0001
India	Fatigue	Tracks 1 and 2	p < .001
		Tracks 1 and 3	p < .002
Lima	Vomiting	Tracks 2 and 3	p < .009
November	Cardiac Counselling	Tracks 1 and 3	p < .038

Note: Track 1 (Blue Team SPs) and Tracks 2 and 3 (Red Team SPs)

cases ranged between 0.30 and 0.39, and the Cronbach's alpha for the remaining three cases were lower than 0.25. There is no significant difference between the coefficients calculated for the complete checklist versus the interaction only checklist (*F* [1, 24} = .520, p < .478). A comparison of the coefficients is located in Table 11.

Item Analysis

A quality assurance strategy recommended by Boulet and colleagues (2003) involves performing an item analysis on the individual checklist items and determining the percentage of examinees receiving or not receiving credit for addressing the item. The checklists (and the accompanying item analysis) can not be presented in order to protect the confidentiality of the cases for future use. A quartile range, assessment of global

Table 11

A comparison of the internal consistency coefficients (Cronbach's alpha) calculated with and without the post-encounter probe (PEP) items

Case	Number of	Number of	Number of	Checklist	Checklist Minus
	Candidate/	PEP	PEP	Coefficient	the PEP items
	SP Items[a]	Questions[b]	Items[c]		Coefficient
Alpha	36	2	7	. 79	.76
Bravo	34	3	7	.64	.49
Charlie	19	2	9	.46	.34
Delta	10	1	3	.40	.49
Echo	30	1	6	.36	.33
Foxtrot	12	1	4	.54	.53
Golf	23	2	10	.36	.24
Hotel	20	3	20	.43	.06
India	23	2	9	.69	.69
Juliet	23	3	12	.52	.58
Kilo	23	3	12	.19	.12
Lima	30	2	6	.65	.61
Mike	23	2	5	.42	.52
November	23	0	0	.54	.54

[a] Number of checklist items comprising the candidate/SP encounter

[b] Number of PEP questions per case

[c] Number of checklist items comprising the PEP

ratings (regarding the pass/fail status of the case), and an assessment of case diagnosis (regarding the pass/fail status of the case), follows.

Quartile Range. A distribution of items for each case was calculated using a quartile range (0 to 24.9%, 25 to 49.9%, 50 to 74.9%, and 75 to 100%). There were several instances where a checklist item was addressed by over half of the candidates. There were also many occasions when a checklist item was not addressed by over half or even over three quarters of the candidates. The quartile distribution for the fourteen cases is presented in Table 12.

Global score and Pass/Fail Status. An investigation was undertaken to ascertain whether a candidate could pass a case, because he or she had addressed enough checklist points to exceed the passing standard, despite having scored a global rating of 1 (poor) or 2 (borderline fail). From a total of 546 candidate/SP interactions, 57 (10.4%) received a global rating of 'poor'. On ten of these occasions the candidate still passed the case. A total of 80 (14.7%) interactions were rated as a 'borderline fail'. On 34 of these interactions, the candidate still passed the case. Correspondingly, there were a total of 157 interactions scored as a 'borderline pass' on the global rating. On 31 of these occasions, the candidate failed the case. Appendix P presents an overview of global score and pass/fail classifications for each case.

Diagnosis and Pass/Fail Status. A third investigation was undertaken to review whether a candidate could fail a case despite an accurate diagnosis of the clinical complaint or pass a case despite an incorrect diagnosis. A review of the four cases with a single diagnosis follows.

In the Alpha case, nine candidates did not correctly diagnose the reason for the

Table 12

The distribution of checklist items addressed by the candidates based on quartile ranges

		Number of Items Addressed (%)			
Case	Number of Checklist Items	0.0 - 24.9	25.0 - 49.9	50.0 - 74.9	75.0 - 100
Alpha	43	5 (11.6%)	7 (16.3%)	14 (32.6%)	17 (39.5%)
Bravo	41	6 (14.6%)	12 (29.3%)	13 (31.7%)	10 (24.4%)
Charlie	28	8 (28.6%)	6 (21.4%)	5 (17.9%)	9 (32.1%)
Delta	13	2 (15.4%)	2 (15.4%)	5 (38.4%)	4 (30.8%)
Echo	36	3 (8.3%)	8 (22.2%)	15 (41.7%)	10 (27.8%)
Foxtrot	16	3 (18.8%)	7 (43.7%)	4 (25.0%)	2 (12.5%)
Golf	33	5 (15.2%)	7 (21.2%)	7 (21.2%)	14 (42.4%)
Hotel	40	11 (27.5%)	4 (10.0%)	15 (37.5%)	10 (25.0%)
India	32	5 (15.6%)	7 (21.9%)	11 (34.4%)	9 (28.1%)
Juliet	35	4 (11.4%)	6 (17.1%)	18 (51.4%)	7 (20.0%)
Kilo	35	5 (14.3%)	5 (14.3%)	10 (28.6%)	15 (42.8%)
Lima	36	2 (5.6%)	10 (27.8%)	13 (36.1%)	11 (30.5%)
Mike	28	2 (7.1%)	5 (17.9%)	12 (42.9%)	9 (32.1%)
November	23	2 (8.7%)	3 (13.0%)	12 (52.2%)	6 (26.1%)

patient's fever, although they were not the same nine that failed the case. Six of the nine candidates that failed the case did not receive credit for the correct diagnosis. It should be noted that despite not getting the correct diagnosis, two of these six candidates were awarded a 'borderline pass' on the global rating. Three of the nine candidates that failed the case did diagnose the source of the fever correctly, but did not accumulate enough points to exceed the minimum performance score for the case. Finally, three candidates who passed the case were unable to make the correct diagnosis.

A total of thirty candidates correctly diagnosed the case (including the three candidates that failed). Six of these thirty candidates passed the case despite scoring a 'borderline fail' on the global score. Incidentally, five of these six 'borderline fail' candidates interacted with the same SP. Although these five 'borderline fail' candidates were able to correctly diagnose the source of the fever and pass the case, the assigned global score could be a reflection of an SP's influence on candidate performance.

In the Bravo case, thirteen candidates were unable to correctly diagnose the case; six of these candidates failed the case and seven passed. Of the seven candidates who passed the case, despite the incorrect diagnosis, three scored a 'poor' and four scored a 'borderline fail' on the global rating. Of the six candidates that failed the case and were unable to get the correct diagnosis, four interacting with one SP (Average Build) and two with the other SP (Obese Build). Of the seven candidates who passed the case (despite the incorrect diagnosis), four interacted with the Obese SP and three interacted with the Average SP.

In the Juliet case, four candidates were unable to make the correct diagnosis, although one of these candidates still passed the case. Fourteen candidates failed the case with eleven receiving credit for the correct diagnosis while three did not receive credit for the diagnosis. The fourth candidate to not get the diagnosis passed the case despite also being awarded a 'borderline fail' by the physician rater. Of the eleven candidates who failed the case, despite making the correct diagnosis, seven were awarded a 'borderline pass' from the physician rater.

In the Mike case, four candidates did not correctly diagnose the case with two candidates passing and two failing the case. The two candidates who failed the case and did not receive credit for the diagnosis both we awarded a 'borderline pass' from the assigned rater. The two candidates who passed the case without making the correct diagnosis were rated 'good' by the assigned rater. Seven candidates failed the case. Six of these candidates correctly diagnosed the problem with four of the candidates being awarded a 'borderline pass' from the assigned rater.

Correlation between an individual case score and mean checklist score

Boulet et al. (2003) recommended Pearson's correlation between each individual case scores and the overall mean scores be calculated to evaluate how well each case is working. Strategy one correlate the case scores (e.g., Alpha) with each candidate's mean score across all the cases (e.g., Alpha through November), while strategy two correlates the case scores (e.g., Alpha) with each candidate's mean score across the remaining cases (e.g., the mean score for Bravo through to November, thus excluding Alpha).

When the scores from each case are compared to the mean score across all cases, only one case (Mike; urinary problem) scored a Pearson's correlation higher than 0.70, three cases scored between .60 and .69, five cases scored between .50 and .59, four cases scored between .40 and .49, and the remaining case had a correlation equal to .30. When the scores from each case are compared using the second strategy (the scores from the case being evaluated are not included in the overall mean score), the resultant coefficient is lower in all fourteen cases. There is a statistically significant difference in the correlations calculated using the two strategies (*F* [1, 26} = 7.166, p < .013). The Pearson's correlation calculated for each case, using both assessment strategies, is located in Table 13.

While there are no negative coefficients to suggest that low ability candidates scored high on a case or high ability candidates scored low on a case, nor do any of the coefficients approach zero, the results indicate that many of the cases are not working particularly effectively, and this is especially so when the coefficients are calculated without the case scores included in the overall mean scores.

Generalizability Analyses

A series of Generalizability analyses were calculated for the checklist, global and total scores using both a one-facet crossed and two-facet nested and assigned designs.

One-facet crossed design. The variance, percentage variance, and Ep^2 for the one-facet crossed design for the checklist, global and total scores are located in Table 14. The percentage of variance accounted for by candidates indicates that there is a difference in the skill level between candidates; findings that are not unexpected. The percentage of variance accounted for by case indicates that there is a difference in the difficulty levels of the cases; again, findings that are not unexpected. The lower percent variance, for candidates and cases, observed in the global score, on all three assessment formats, could be an artefact of the smaller assessment scale used (the five point rating scale in comparison to a multi-item checklist) or it might illustrate the limitation of one global

Table 13

Correlation (Pearson's r) between case scores and overall mean case scores

Case Name	Case Problem	Comparison across all cases	Comparison minus the specific case
Alpha	Fever	0.65**	0.55**
Bravo	Headache	0.62**	0.50**
Charlie	Infection	0.30	0.19
Delta	Informed Consent	0.51**	0 .37*
Echo	Risk Assessment	0.42**	0.32*
Foxtrot	Metastasis of cancer	0.51**	0.36*
Golf	Shortness of Breath	0.58**	0.51**
Hotel	Flu Symptoms	0.44**	0.36**
India	Fatigue	0.47**	0.34*
Juliet	Hand Problem	0.56**	0.47**
Kilo	Personality Changes	0.60**	0.26
Lima	Vomiting	0.46**	0.34*
Mike	Urinary Tract	0.77**	0.71**
November	Cardiac Counselling	0.55**	0.45**

* indicates that the correlation is significant at the 0.05 level (2-tailed)

** indicates that the correlation is significant at the 0.01 level (2-tailed).

Table 14

The variance, percentage variance and generalizability coefficients for the one-facet crossed design (candidates x cases)

Assessment Format	Facets	Variance	% Variance	Ep^2
Checklist score	Candidates(p)	28.6617	18.8%	0.80 (0.795)
	Cases (c)	20.2982	13.3%	
	p x c, residual	103.4260	67.9%	
	Total	152.3859	100%	
Global score	Candidates (p)	0.1353	11.7%	0.67 (0.674)
	Cases (c)	0.1029	8.9%	
	p x c, residual	0.9155	79.4%	
	Total	1.1537	100%	
Total score	Candidates(p)	31.0242	18.4%	0.79 (0.789)
	Cases (c)	21.4590	12.8%	
	p x c, residual	115.8714	68.8%	
	Total	168.3546	100%	

Note: Candidates were designated as 'participants' (p) to differentiate from cases (c) in the G_String_II programming

scale to evaluate performance.

The percentage of variance accounted for by candidates by cases (interaction) suggests there is a difference in how the candidates managed the cases (e.g., a candidate might find one case easier to manage than another). It should also be noted that the variance as a result of the interaction effect (candidates by cases) and the random error can not be separated and are combined together with the residual.

The reliability coefficient (Ep$^{2)}$ calculated for all three assessment scores do not meet the criteria set for high-stakes assessment. The lower Ep^2 for the global score indicates the global score is being used less reliably for the evaluation of performance compared to the checklist score, while the lower total score Ep^2 indicates that the global score is influencing the reliability coefficient.

Two-facet Nested Design (SPs nested into Cases). The variance, percentage variance, and Ep^2 for the two-facet nested design (based on the three assessment tracks) for the checklist, global and total scores is located in Table 15.

Based on the calculations between the three examination tracks, the interpretations in the percentage of variance accounted for by candidates, cases, and candidates by cases are similar to those in the one-facet design. There are differences in skill level between the candidates, there is a difference in case difficulty, and the interaction effect suggests there are differences in how the candidates managed the cases. For a nested design, a non-zero variance component for SP nested into a case indicates that there are differences in the candidates' scores due to the SPs selected to portray the case (Boulet et al., 2003). Based on the calculations between tracks, 17.5%, 23.1%, and 20.2% of the variance (checklist, global, and total score, respectively) is accounted for by the SPs between the three tracks. The reliability coefficient calculated for all three assessment scores do not meet the criteria set for high-stakes assessment, The Ep^2 for the global score is lower than both the checklist and total scores, which could suggest the global score is less reliable for evaluating performance than the checklist score.

Table 15

The variance, percentage variance and generalizability coefficients (based on the three assessment tracks) for the two-facet nested design (SP nested into case)

Assessment Format	Facets	Variance	% Variance	Ep^2
Checklist score	Candidates(p)	30.9609	20.0%	0.84 (0.836)
	Cases (c)	11.6028	7.5%	
	Cases:SP (c:sp)	27.0727	17.5%	
	p x c:sp, residual	85.0487	55.0%	
	Total	154.6851	100%	
Global score	Candidates(p)	0.1474	12.6%	0.74 (0.738)
	Cases (c)	0.0162	1.4%	
	Cases:SP (c:sp)	0.2698	23.1%	
	p x c:sp, residual	0.7323	62.8%	
	Total	1.1657	100%	
Total score	Candidates(p)	34.3664	20.0%	0.84 (0.839)
	Cases (c)	10.2961	6.0%	
	Cases:SP (c:sp)	34.7457	20.2%	
	p x c:sp, residual	92.2903	53.8%	
	Total	171.6985	100%	

Note: Candidates were designated as 'participants' (p) to differentiate from cases (c) in the G_String_II programming

The variance, percentage variance, and Ep^2 for the two-facet nested design (based on the two SP teams) for the checklist, global and total scores are located in Table 16.The interpretations in the percentage of variance accounted for by candidates, cases, and candidates by cases, based on the two teams of SPs, are similar to those in the one-facet design and the two-facet nested. There are differences in skill level between the candidates, there is a difference in case difficulty, and the interaction effect indicates there are differences in how the candidates managed the cases.

As with the two-facet nested comparison between the three assessment tracks, a non-zero variance component was observed for the SPs nested into a case, again, indicating that there are differences in the candidates' scores due to the SPs selected to portray the case (Boulet et al., 2003). Based on the calculations between SPs, 14.3%, 20.4%, and 16.7% of the variance (checklist, global, and total score, respectively) is accounted for between the SPs.

The reliability coefficient calculated for all three assessment scores do not meet the criteria for set high-stakes assessment The EP^2 for the global score is lower than both the checklist and total scores, which could suggest the global score is less reliable for evaluating performance than the checklist score.

Table 16

The variance, percentage variance and generalizability coefficients (based on the two SP teams) for the two-facet nested design (SP nested into case)

Design (Score)	Facets	Variance	% Variance	Ep^2
Checklist score	Candidates(p)	30.1011	19.6%	0.82 (0.819)
	Cases (c)	8.2851	5.4%	
	Cases:SP (c:sp)	22.0887	14.3%	
	p x c:sp, residual	93.3504	60.7%	
	Total	153.8253	100%	
Global score	Candidates(p)	0.1406	11.8%	0.71 (0.710)
	Cases (c)	0.0000	0%	
	Cases:SP (c:sp)	0.2423	20.4%	
	p x c:sp, residual	0.8050	67.8%	
	Total	1.1879	100%	
Total score	Candidates(p)	32.9716	19.4%	0.82 (0.818)
	Cases (c)	5.9495	3.5%	
	Cases:SP (c:sp)	28.5119	16.7%	
	p x c:sp, residual	102.8707	60.4%	
	Total	170.3037	100%	

Note: Candidates were designated as 'participants' (p) to differentiate from cases (c) in the G_String_II programming

Two Facet Design (Physician raters). The variance, percentage variance, and Ep^2 for the two-facet nested design (based on physician rater assignment) for the checklist, global and total scores is located in Table 17. The interpretations in the percentage of variance accounted for by candidates, cases, and candidates by cases, based on physician rater assignment, are similar to those in the one-facet design and the two-facet nested (SPs). There are differences in skill level between the candidates, there is a difference in case difficulty, and the interaction effect suggests there are differences in how the candidates managed the cases.

As with the two-facet nested comparison between the three assessment tracks, a non-zero variance component was observed for the physician raters assigned to a case. This indicates that there are differences in the candidates' scores due to the physician raters assessing the cases. Based on the calculations, 5.4%, 8.4%, and 7.5% of the variance (checklist, global, and total score, respectively) is accounted for by the physician raters assigned to a case.

A non-zero variance is also accounted for by case crossed by rater assigned to cases. Based on the calculations, 9.9%, 14.0%, and 10.2% of the variance (checklist, global, and total score, respectively) is accounted for by the physician raters. These results are indicative of the variance associated with the ten physician raters from the morning session being assigned to a different case for the afternoon track. This might suggest that the time between the morning and afternoon session was not long enough for the physician raters to prepare themselves for a new case or it might indicate that the physician rater newly assigned to the afternoon session may have had a different approach to the assessment of the case than the raters from the morning sessions.

Table 17

The variance, percentage variance and generalizability coefficients for the two-facet design based on physician rater assigned to a case or cases

Assessment Format	Facets	Variance	% Variance	Ep^2
Checklist score	Candidates(p)	30.324	19.6%	0.83 (0.832)
	Cases (c)	16.058	10.4%	
	Raters (e) case assignment	8.398	5.4%	
	p x c(e), residual	84.962	54.8%	
	Case x Rater (assignment)	15.36	9.9%	
	Total	155.102	100%	
Global score	Candidates(p)	0.145	12.5%	0.74 (0.736)
	Cases (c)	0.023	2.0%	
	Raters (e) case assignment	0.097	8.4%	
	p x c(e), residual	0.731	63.1%	
	Case x Rater (assignment)	0.162	14.0%	
	Total	1.158	100%	
Total score	Candidates(p)	33.557	19.6%	0.84 (0.836)
	Cases (c)	15.468	9.0%	
	Raters (e) case assignment	12.921	7.5%	
	p x c(e), residual	92.167	53.7%	
	Case x Rater (assignment)	17.486	10.2%	
	Total	171.599	100%	

The reliability coefficient calculated for all three assessment scores do not meet the criteria for high-stakes assessment. The EP^2 for the global score is lower than both the checklist and total scores, which could suggest the global score is less reliable for evaluating performance than the checklist score.

Two Facet Design (Physician rater and SP combinations). The variance, percentage variance, and Ep^2 for the two-facet nested design (based on the Rater/SP assignment) for the checklist, global and total scores is located in Table 18.

The interpretations in the percentage of variance accounted for by candidates, cases, and candidates by cases are similar to those in the one-facet design and the two-facet nested. There are differences in skill level between the candidates, there is a difference in case difficulty, and the interaction effect suggests there are differences in how the candidates managed the cases.

As with the two-facet nested comparison between the three assessment tracks, a non-zero variance component was observed for the rater/SPs combinations assigned to a case indicating that there are differences in the candidates' scores due to these assignments. Based on the calculations, 16.8%, 22.2%, and 18.6% of the variance (checklist, global, and total score, respectively) is accounted for by the assigned combinations. This could indicate that the physician raters scoring of the candidate's performance might be influenced by the portrayal by the SP. The reliability coefficient calculated for all three assessment scores do not meet the criteria set for high-stakes assessment. The Ep^2 for the global score is lower than both the checklist and total scores, which could suggest the global score is less reliable for evaluating performance than the checklist score.

Table 18

The variance, percentage variance and generalizability coefficients for the two-facet design based on physician rater/SP assignments

Assessment Format	Facets	Variance	% Variance	Ep^2
Checklist score	Candidates(p)	30.554	20.5%	0.84 (0.842)
	Cases (c)	13.629	9.1%	
	Rater/SP combinations	25.110	16.8%	
	p x c, residual	80.080	53.6%	
	Total	149.373	100%	
Global score	Candidates(p)	0.145	12.5%	0.74 (0.735)
	Cases (c)	0.026	2.2%	
	Rater/SP combinations	0.258	22.2%	
	p x c, residual	0.732	63.0%	
	Total	1.161	100%	
Total score	Candidates(p)	33.794	19.8%	0.84 (0.837)
	Cases (c)	13.020	7.6%	
	Rater/SP combinations	31.614	18.6%	
	p x c, residual	92.326	54.1%	
	Total	170.754	100%	

Note: Candidates were designated as 'participants' (p) to differentiate from cases (c) in the G_String_II programming

Communication Checklist

The physician rater and the SP rated the communication skills of the candidates using a 13-item checklist. A copy of the communication checklist used by the physician raters is located in Appendix Q.

Cronbach's alpha. The overall internal consistency, for the fourteen cases, based on the physician raters scores equaled 0.92 (0.918), while the overall internal consistency for the SP raters scores equaled 0.92 (0.921). The internal consistency for the physician raters, by case, ranged between 0.77 and 0.97. The internal consistency for the SP raters, by case, ranged between 0.58 and 0.94. Differences between the physician rater and SP (coefficients larger than 0.10) were noted on the Bravo (headache), Charlie (infection, Golf (shortness of breath), and Lima (vomiting) cases. A comparison between the physician rater and SP mean, standard deviation, and internal consistency across the fourteen cases is located in Table 19.

The internal consistency for the thirteen communication items is generally lower for both the physician raters and SPs. The internal consistency for the physician raters ranged between 0.16 and 0.56, while the internal consistency for the SP raters ranged between 0.07 and 0.53. Differences between the physician rater and SP (coefficients larger than 0.10) were noted on item nine (the doctor let the patient express ideas in planning treatment, tests, or follow-up), item eleven (the doctor used non-technical language), item twelve (the doctor was careful and thorough), and item thirteen (the patient was satisfied with medical care received). A comparison between the physician rater and SP mean, standard deviation, and internal consistency across each item on the checklist (inter-item) is located in Table 20.

Table 19

A comparison between the physician rater and SP communication scores (mean, standard deviation, and internal consistency) for each case

Case	Mean(SD)		Cronbach's alpha	
	Physician	SP	Physician	SP
Alpha (Fever)	3.4 (0.76)	3.7 (0.56)	0.93	0.88
Bravo (Headache)	3.8 (0.76)	4.3 (0.45)	0.91	0.79
Charlie (Infection)	3.8 (0.33)	4.3 (0.50)	0.77	0.89
Delta (Informed Consent)	3.9 (0.84)	3.9 (0.58)	0.97	0.90
Echo (Risk Assessment)	3.8 (0.45)	3.8 (0.52)	0.89	0.92
Foxtrot (Metastasis Cancer)	3.8 (0.59)	3.7 (0.44)	0.88	0.89
Golf (Shortness of Breath)	3.4 (0.44)	3.4 (0.21)	0.86	0.58
Hotel (Flu Symptoms)	3.9 (0.60)	4.0 (0.69)	0.87	0.90
India (Fatigue)	3.6 (0.82)	3.7 (0.80)	0.93	0.94
Juliet (Hand Problem)	3.5 (0.56)	3.3 (0.61)	0.87	0.87
Kilo (Personality Change)	3.8 (0.35)	4.0 (0.52)	0.82	0.88
Lima (Vomiting)	3.4 (0.45)	3.4 (0.93)	0.78	0.93
Mike (Urinary Tract)	4.0 (0.43)	4.4 (0.46)	0.89	0.91
November (Cardiac Counselling)	3.6 (0.54)	3.9 (0.39)	0.87	0.89

Table 20

A comparison between the physician rater and SP communication scores (mean, standard deviation, and internal consistency) for each checklist item

Question	Mean(SD)		Cronbach's alpha	
	PE	SP	PR	SP
Understand how the patient saw things	3.7 (0.29)	3.9 (0.24)	0.42	0.36
Sensed what the patient was feeling	3.7 (0.30)	3.9 (0.28)	0.49	0.53
Doctor took no notice of what patient felt*	3.9 (0.27)	3.9 (0.28)	0.16	0.26
Doctor's response was fixed and automatic*	3.9 (0.35)	4.0 (0.32)	0.48	0.52
Patient treated with respect and courtesy	4.2 (0.16)	4.4 (0.16)	0.25	0.07
Patient was able to explain problem to doctor	3.6 (0.27)	3.9 (0.24)	0.26	0.34
The doctor explained what might be wrong	3.6 (0.28)	3.7 (0.33)	0.29	0.39
The doctor explained treatment and tests	3.5 (0.29)	3.7 (0.33)	0.32	0.33
The doctor allowed patient input on treatment	3.2 (0.38)	3.2 (0.33)	0.51	0.37
Doctor gave patient a chance to ask questions	3.4 (0.38)	3.6 (0.36)	0.55	0.49
The doctor used non-technical language	3.9 (0.27)	4.0 (0.29)	0.53	0.38
The doctor was careful and thorough	3.6 (0.36)	4.0 (0.26)	0.56	0.41
The patient feels satisfied with cares	3.4 (0.32)	3.8 (0.31)	0.53	0.43

Note: PR = Physician Rater; SP = Standardized Patient

* questions are reverse coded

Generalizability analyses (SP raters). Generalizability analyses were calculated for each communication checklist item (Item 1 to Item 13). Non-zero variance components calculated with the SPs' data were observed for candidates (p), cases (c), Candidates crossed with cases (p x c), and SPs nested into case (sp:c). The complete generalizability results for each checklist item are located in Appendix R. A brief summary of the percent variance follows.

Variance due to candidates (p) ranged from 1.9% to 7.0%, indicating there are differences between the candidates on their ability to communicate, as scored by the SP raters. A percent variance higher than 6.0% was observed on three items, including: Item 2 (6.3%) (the doctor usually sensed or realized what I was feeling), Item 4 (7.0%) (the doctor's response was usually so fixed and automatic that I didn't really get through to him/her), and Item 13 (6.9%) (I feel satisfied with the medical care I received). The higher variance results on items 2 and 4 might be a reflection of the candidates' English as a second language skills, while the variance associated with the patient satisfaction item could reflect different expectations of performance by the candidates on the part of the SPs (trained for expectations or personal expectations).

Variance due to cases (c) ranged from zero to 23.2%, indicating there are differences between how the items were evaluated across the fourteen cases. Zero variance was calculated on two items, including: Item 5 (the doctor treated me with respect and courtesy) and Item 11 (the doctor used understandable and non-technical language). A percent variance higher than 10% was noted on seven checklist items, including: Item 1 (23.2%) (the doctor wanted to understand how I saw things), Item 2 (11.3%) (the doctor usually sensed or realized what I was feeling), Item 6 (14.9%) (I was able to explain my problem to the doctor as fully needed, Item 8 (12.7%) (the doctor explained what treatment, tests, or other follow-up is going to happen), Item 9 (17.3%) (the doctor gave me an opportunity to express my feelings or ideas in planning treatment, tests, or follow-up), Item 10 (11.7%) (the doctor gave me the opportunity to ask questions, and Item 12 (11.4%) (the doctor was careful and thorough). The higher variance on items 8 and 9, for example, could be a reflection of the instructions presented to the candidates prior to entering the examination room. For many of the cases, the candidates were instructed to take a history and perform a focused physical assessment. No direction was provided about outlining test or treatments to the SP or soliciting the SP's opinion on treatment or testing protocols beyond "close the encounter appropriately" (the Golf [shortness of breath] case). Higher variance results could be due to the SP expecting the candidates to explain or ask for input on treatment plans, and when these aspects were not forthcoming the candidate(s) might not have received credit on a communication item they were unaware they should be addressing or the SP simply circled 3 (not sure) as a 'not applicable to this case' option was not provided.

Candidate by case (p x c) variance ranged from 51.3% to 71.1%, which suggests there are differences in the candidates' communication scores across the fourteen cases. A percent variance higher than 60% was noted on six checklist items, including: Item 3 (71.1%) (the doctor just took no notice of some of the things I thought or felt), Item 4 (70.0%) (the doctor's response was usually so fixed and automatic that I didn't really get through to him/her), Item 5 (62.8%) (the doctor treated me with respect and courtesy), Item 7 (68.6%) (the doctor explained what might be the matter with me), Item 8 (67.1%) (the doctor explained what treatment, tests, or other follow-up is going to happen), and

Item 12 (60.3%) (the doctor was careful and thorough). Higher variance on the p x c facet for six of the checklist items might also be a reflection of the instructions provided to the candidates prior to entering the examination room in addition to the expectations of the SPs based on their training for the case.

Variance accounted for by the SPs nested in the cases ranged from 17.7% to 37.5% , indicating that there are differences in the SPs' assessment of communication skills across cases. A percent variance higher than 20% were noted on ten items, including: Item 1 (21.2%) (the doctor wanted to understand how I saw things), Item 2 (27.0%) (the doctor usually sensed or realized what I was feeling), Item 5 (33.8%) (the doctor treated me with respect and courtesy), Item 6 (29.2%) (I was able to explain my problem to the doctor as fully needed, Item 7 (26.1%) (the doctor explained what might be the matter with me), Item 9 (27.3%) (the doctor gave me an opportunity to express my feelings or ideas in planning treatment, tests, or follow-up), Item 10 (27.2%) (the doctor gave me the opportunity to ask questions, Item 11 (36.5%) (the doctor used understandable and non-technical language), Item 12 (22.6%) (the doctor was careful and thorough), and Item 13 (37.5%) (I was satisfied with the medical care I received).

Generalizability analyses (Physician rater scores). Generalizability analyses were calculated for each communication checklist item (Item 1 to Item 13). Non-zero variance components calculated with the physician raters data were observed for candidates (p), cases (c), Candidates crossed with cases (p x c), physician raters assigned to case (r in c), and case crossed by rater assigned to case (c x r in c).

Variance due to candidates (p) ranged from 1.2% to 9.1%, indicating there are differences between the candidates ability to communicate based on the specific communication item. A percent variance higher than 6.0% was observed for five items, including: Item 2 (6.5%) (the doctor usually sensed or realized what the patient was feeling), Item 4 (6.6%) (the doctor's response was usually so fixed and automatic that the patient did not get through to him/her), Item 11 (7.9%) (the doctor used non-technical language), Item 12 (8.8%) (the doctor was careful and thorough), and Item 13 (9.1%) (the patient feels satisfied with the medical care received). A comparison between the SP and physician raters' communication items with a percent variance higher than 6.0% can be found in Appendix R.

Variance due to cases (c) ranged from zero to 9.2%, indicating there are differences between how the items were evaluated across the cases. Zero variance between cases was calculated for nine of the thirteen communication checklist items (compared to two items for the SP raters; items 5 [respect and courtesy] and 11 [used understandable and non-technical language]). A percent variance above 10% was not observed on any of the communication items (compared to seven items for the SP raters). A comparison between the SP and physician raters' communication items with a percent variance higher than 10.0% can be found in Appendix R.

Candidate by case (p x c) variance ranged from 44.4% to 86.1%, which suggests there are differences in the candidates' communication scores across the fourteen cases. These findings are not unexpected as the results typically indicate the differences in how the candidates managed the cases (e.g., counselling cases versus clinical competency cases). A percentage variance higher than 60% was noted on five items, including: Item 6 (70.6%) (the patient was able to explain his/her problem to the doctor as fully needed), Item 7 (86.1%) (the doctor explained things to the patient so that they know what may be the matter with them), Item 8 (83.8%) (the doctor explained what treatment, tests, or other follow-up is going to happen), Item 10 (60.3%) (the doctor gave the patient the opportunity to ask questions), and Item 11 (62.6%) (the doctor used understandable and non-technical language. Two of these five items were also noted in the SP results (items 7 [what is the matter] and 8 [the doctor explained treatment, etc]). A comparison between the SP and physician raters' communication items with a percent variance higher than 60.0% can be found in Appendix S.

Variance accounted for by the physician raters assigned within the cases ranged from 3.0% to 51.4% , which indicates that there are differences in the physician raters' assessment of communication skills between the raters assigned to a case. A percent variance higher than 20% were noted in ten items, including: Item 1 (30.1%) (the doctor wanted to understand how the patient saw things), Item 2 (28.0%) (the doctor usually sensed or realized what the patient was feeling), Item 3 (35.5%) (the doctor just took no notice of some things that the patient thought or felt), Item 4 (29.2%) (the doctor's response to the patient was usually so fixed and automatic that the patient didn't really get through to him/her), Item 5 (51.4%) (the doctor treated the patient with respect and courtesy), Item 6 (24.7%) (the patient was able to explain the problem to the doctor as fully needed, Item 9 (32.7%) (the doctor gave the patient an opportunity to express his/her feelings or ideas in planning treatment, tests, or follow-up), Item 10 (24.7%) (the doctor gave the patient the opportunity to ask questions, Item 11 (29.5.5%) (the doctor used understandable and non-technical language), and Item 13 (25.8%) (the patient was satisfied with the medical care received). A comparison between the nested SP and assigned physician raters' communication items with a percent variance higher than 20.0% can be found in Appendix T.

ANOVA (SP raters). A comparison in the communication scores, for the fourteen cases, between the three assessment tracks was evaluated using an ANOVA. Statistically significant differences ($p < .05$), based on the three assessment tracks, were found on nine cases (Charlie [infection], Delta [informed consent], Golf [shortness of breath], Hotel [flu symptoms], India [fatigue], Juliet [hand problem], Lima [vomiting], Mike [urinary tract] and November [cardiac counselling]). The ANOVA results across the fourteen cases, based on the three assessment tracks, are located in Appendix U.

A comparison in the communication scores, for the thirteen checklist items, between the three assessment tracks was also evaluated with an ANOVA. Statistically significant differences ($p < .05$), based on the three assessment tracks, were found on four checklist items (item 3 [the doctor took no notice of some things I thought or felt], item 6 [I was able to explain the problem to the doctor], item 7 [the doctor explained things so I knew what might be the matter with me], and item 10 [the doctor gave me the opportunity to ask questions]). The ANOVA results comparing the communication skills scores on the thirteen items, based on the three assessment tracks, can be reviewed in Appendix V.

A comparison in the communication scores, for the fourteen cases, between the SPs assigned to each case was evaluated using an ANOVA. There was a statistically significant difference ($p < .05$) between the two SPs on nine cases (Charlie [infection], Delta [informed consent], Golf [shortness of breath], Hotel [flu symptoms], India [fatigue], Juliet [hand problem], Kilo [personality changes], Lima [vomiting], and Mike [urinary tract]). The ANOVA results across the fourteen cases, based on the two SPs, are located in Appendix U.

A comparison in the communication scores, for the thirteen checklist items, between the two SP teams was evaluated using an ANOVA. A statistically significant difference (p < .05) between the two SP teams was found on two checklist items (item 6 [I was able to explain the problem to the doctor] and item 7 [the doctor explained what was wrong with me]). The ANOVA results comparing the communication skills scores on the thirteen items, based on the two SP teams, can be reviewed in Appendix V.

ANOVA (physician raters). As with the SP communication skills data, an ANOVA was performed to compare the assigned physician raters scores on the cases and checklist items. There was a statistically significant difference (p < .05) between the on three assessment tracks on nine cases (Alpha [fever], Bravo [headache], Delta [informed consent], Echo [risk assessment], Golf [shortness of breath], Hotel [flu symptoms], India [fatigue], Kilo [personality changes], and Lima [vomiting]). The ANOVA results across the fourteen cases, based on the three teams of physician raters, are located in Appendix W.

There was a statistically significant difference (p < .05) between the physician rater teams on four checklist items (Item 1 [the doctor wanted to understand how the patient saw things], Item 6 [the patient was able to explain the problem to the doctor], Item 9 [the doctor gave the patient the opportunity to express his/her feelings or ideas in planning treatment, tests, or follow-up], and Item 10 [the doctor gave the patient the opportunity to ask questions]). The ANOVA results across the thirteen checklist items, based on the three teams of physician raters, are located in Appendix X.

Summary Study One

The descriptive statistics revealed an overall mean score (using the total score data), across all fourteen cases, of 59%. ANOVA results demonstrated statistically significant differences on several cases between assessment tracks and SPs. Internal consistency measures (Cronbach's alpha) indicated low coefficient scores on many of the cases with an item analysis revealing that many checklist items were not addressed by the candidates. A relationship between case and overall scores (not including the case scores) showed a low to moderate correlation (Pearson's *r*) suggesting that lower skilled candidates were performing reasonably well on some cases where the higher skilled candidates were not performing well on those same cases. Generalizability analyses demonstrated that approximately one-fifth of the variance could be attributed to differences in SP portrayal, while almost one-fifth of the variance could be ascribed to differences in physician rating.

The communication skills assessment revealed moderate to high internal consistency for each case. Low to moderate internal consistency was calculated for each checklist item across cases (inter-item). The percent variance for each checklist item on participants, cases, p x c, and for the nested SPs or assigned physician raters varies considerably from item to item, suggesting there are differences in how the items are being rated based on the requirements of the case and there is considerable differences between the nested SPs and assigned physician raters on the scoring of the communication skills items.

Study Two

Fifteen candidates from the WAAIP OSCE were randomly selected and their videotaped performances rated by one physician and two non-physician raters. The objective was to establish inter-rater reliability between the two presiding raters in the WAAIP OSCE (physician rater and SP) versus three videotape raters (another physician and two non-physicians). The WAAIP OSCE was comprised of 14 cases. During the course of computer scanning, hardcopy data from one of the non-physician raters was lost for the Foxtrot case. The statistical analyses will be performed between scores collected from the physician raters and the remaining non-physician rater for this case.

Demographics

Age and Gender. Seven males (46.7%) and eight females (53.3%) were randomly selected. There is a significant difference between age based on gender (*F* [1, 13] = 5.048, p < .043). Summary statistics for age and gender are located in Table 21.

Table 21

Descriptive statistics for age (in years) and gender of the candidates (N = 15)

Category	Number	M	SD	Min.	Max.	Range
Male	7	46.0	8.0	34	55	21
Female	8	38.3	5.3	29	45	16
Total	15	41.9	7.6	29	55	26

Country of Origin. Ten countries of origin are represented in the sample. While three of the countries are represented by multiple participants; most are represented by one. Due to the risk of identifying a participant based on country of origin and date of the

WAAIP examination, data for general regions will be presented. Three (20.0%) candidates are originally from Asia, two (13.3%) are from Eastern Europe, nine (60.0%) originate from the Middle East. The remaining participant (6.7%) will be categorized as 'Other" to avoid potential identification.

Years Since Completing Medical School Training. All of the candidates have an MD degree. There is a significant difference between the number of years since medical school graduation based on gender (*F* [1, 13] = 6.353, p < 0.026).

Internship and Residency Training. Of the fifteen participants, nine (60.0%) have completed an Internship and one had not (6.7%). Internship information was not provided by five candidates (33.3%). Internship completion rates, based on gender, are located in Table 22. Of the nine candidates who have completed an Internship, four (44.4%), all male, have completed a Residency program.

Table 22

Internship program completion rates based on gender

Gender	Total	Completed Internship	Not Completed Internship	Information Not Provided
Male	7	6	N/A	1
Female	8	3	1	4

OSCE Results

The physician raters who evaluated the candidates during the course of the WAAIP OSCE will be identified collectively as WP (WAAIP physician). The videotape physician rater will be designated as VP, while the two video non-physician raters will be identified as NP1 and NP2. Significant differences between the four raters are noted on the Bravo (*F* [3, 56} = 4.501, p < .007), Delta (*F* [3, 56} = 14.745, p < .001), and Foxtrot (*F* [3, 56} = 4.944, p < .004) cases. A summary comparing the four raters' total scores (mean and standard deviation) on the fourteen OSCE cases are presented in Table 23. Descriptive statistics comparing the physician raters (WP and VP) and non-physician raters (NP1 and NP2) across the three assessment formats is presented in Table 24.

Analysis of Variance (ANOVA) – Physician raters versus Non-Physician raters

Checklist score. There was a significant difference between the physician and non-physician raters on four of the fourteen cases: Bravo (*F* [1, 58} = 9.092, p < .004), Delta (*F* [1, 58} = 8.697, p < .005), Echo (*F* [1, 58} = 5.314, p < .025), and Golf (*F* [1, 58} = 5.040, p < .029).

Global score. There was a significant difference between the physician and non-physician raters on three of the fourteen cases: Delta (*F* [1, 58} = 9.800, p < .003), Golf (*F* [1, 58} = 4.698, p < .034), and India (*F* [1, 58} = 4.584, p < .036).

Total score. There was a significant difference between the physician and non-physician raters on three of the fourteen cases: Bravo (*F* [1, 58} = 6.541, p < .013), Delta (*F* [1, 58} = 10.743, p < .002), and Golf (*F* [1, 58} = 5.418, p < .023).

Inter-Rater Reliability (Intra-Class Coefficient)

The ICC on the checklist scores range from 0.06 to 0.87 for the physician raters (PR) and 0.61 to 0.95 for the non-physician raters (NP). Global ICC results range from zero to 0.63 for the PR and 0.03 to 0.67 for the NP. The ICC on the total scores range from 0.04 to 0.82 for the PR and 0.54 to 0.94 for the NP. These results reflect the pooling of scores

Table 23

Summary statistics for the fourteen OSCE cases comparing the total score from the WAAIP physician raters (WP), the videotape physician (VP) and two non-physician videotape raters (NP1 and NP2) for the 15 randomly selected participants

OSCE Case	Mean(SD)	Mean(SD)	Mean(SD)	Mean(SD)
(Maximum Score)	(WP)	(VP)	(NP1)	(NP2)
Alpha (48)	30.1 (4.7)	29.4 (4.0)	28.2 (4.5)	28.9 (3.9)
Bravo (46)	30.9 (5.4)*	28.1 (4.4)	27.9 (4.7)	24.9 (3.2)*
Charlie (22)	16.7 (2.7)	16.1 (3.2)	15.5 (3.7)	14.9 (2.6)
Delta (18)	12.1 (1.9)*	7.9 (2.3)*	7.9 (2.3)*	7.7 (1.9)*
Echo (41)	25.1 (3.2)	24.4 (4.3)	23.3 (5.1)	22.2 (3.6)
Foxtrot (21)	11.7 (2.8)	10.5 (2.6)	11.5 (3.1)	N/A
Golf (38)*	23.9 (4.0)*	21.1 (2.6)	21.4 (2.8)	19.7 (2.9)*
Hotel (45)	24.9 (3.5)	23.5 (4.3)	24.6 (4.3)	23.9 (4.0)
India (37)	21.1 (3.6)	19.5 (2.6)	18.6 (2.7)	19.8 (2.4)
Juliet (40)	24.3 (4.9)	25.0 (4.0)	25.2 (5.0)	24.9 (3.5)
Kilo (40)	26.3 (2.8)	26.6 (3.0)	27.5 (3.1)	26.3 (2.0)
Lima (41)	25.1 (5.5)	23.4 (5.6)	23.4 (6.8)	22.1 (4.5)
Mike (33)	22.6 (3.3)	21.3 (2.8)	21.9 (3.4)	21.6 (2.6)
November (28)	17.8 (2.6)	17.9 (3.5)	19.3 (2.9)	18.1 (2.4)

Note: * Statistical differences between raters were noted for the following cases: Bravo (WP and NP2); Delta (WP and all Videotape raters); Golf (WP and NP2).

Table 24

Descriptive statistics comparing the physician and non-physician scores

Case	Checklist		Global		Total	
	PE	NP	PE	NP	PE	NP
Alpha	26.8 (3.6)	25.9 (3.8)	3.0 (1.0)	2.6 (0.9)	29.7 (4.3)	28.6 (4.1)
Bravo	26.3 (4.3)	23.2 (3.5)	3.2 (1.0)	3.2 (0.9)	29.5 (5.0)	26.4 (4.2)
Charlie	14.3 (2.3)	13.9 (2.7)	2.1 (1.0)	1.7 (0.8)	16.3 (3.0)	15.3 (3.2)
Delta	7.1 (1.7)	5.9 (1.5)	2.9 (1.3)	1.9 (1.0)	10.0 (2.9)	7.7 (2.1)
Echo	21.7 (3.1)	19.7 (3.7)	3.0 (0.8)	3.1 (0.9)	24.7 (3.8)	22.8 (4.4)
Foxtrot	8.7 (2.1)	8.1 (2.3)$^{\phi}$	2.9 (1.3)	3.4 (0.9)$^{\phi}$	11.6 (2.8)	11.5 (3.1)$^{\phi}$
Golf	19.5 (2.8)	18.0 (2.4)	3.0 (0.9)	2.6 (0.7)	22.5 (3.6)	20.5 (2.9)
Hotel	21.5 (3.2)	21.4 (3.)	2.7 (1.2)	2.8 (0.9)	24.2 (3.9)	24.3 (4.1)
India	17.3 (2.7)	17.1 (2.0)	2.5 (0.9)	2.1 (0.8)	20.3 (3.2)	19.2 (2.6)
Juliet	21.4 (3.6)	22.1 (3.4)	3.2 (1.0)	3.0 (1.0)	24.6 (4.4)	25.1 (4.3)
Kilo	23.0 (2.6)	23.5 (2.3)	3.4 (0.6)	3.4 (0.6)	26.4 (2.9)	26.9 (2.6)
Lima	21.4 (4.7)	19.9 (4.8)	2.9 (1.1)	2.9 (1.0)	24.3 (5.6)	22.7 (5.7)
Mike	18.4 (2.6)	18.2 (2.5)	3.6 (0.6)	3.5 (0.6)	21.9 (3.1)	21.7 (3.0)
November	14.5 (2.5)	15.4 (2.1)	3.4(1.0)	3.3 (0.7)	17.9 (3.0)	18.7 (2.7)

Note: PE (Physician Examiner); NP (Non-Physician Examiner)

$^{\phi}$ Data missing for one non-physician rater between the WP and VP rater, while the NP results demonstrate the relative consistency of scores due to the standardized training they received.

The ICC results for some cases are moderate to high on the checklist and total scores assessment format but are generally lower for the global scale, which indicates less reliability when using the global format for both the physician and non-physician raters. The ICC results comparing physician and non-physician raters are presented in Table 25.

Generalizability Analyses

A series of Generalizability analyses were performed, for the physician and non-physician data, using a two-facet crossed design (Candidates x Cases x Raters).

Physician raters. Notable observations include the non-zero variance in the rater facet, which indicates differences between the WP and VP raters. This is especially apparent on the global score (9.8%) but less so on the checklist score (2.7%). This is likely a reflection of the checklist format that provides more structure to the evaluation, in comparison to the global score, which could be influenced by subjective interpretations of what constitutes appropriate assessment and management. Zero variance is accounted for the participant x rater interaction, demonstrating there are no differences in how the physician raters evaluated the candidates. A non-zero variance is observed in the case x rater interaction, which indicates there are differences in how the physician raters are evaluating the cases. The variance component on the p x c x r, residual interaction, is likely indicative of the small number of candidates selected for the study (fifteen from the original 39). The variance components for the physician raters are located in Table 26.

Table 25

ICC comparing the physician and non-physician scores for each case

Case	Checklist		Global		Total	
	PR	NP	PR	NP	PR	NP
Alpha	0.52	0.68	0.21	0.33	0.57	0.70
Bravo	0.61	0.62	0.61	0.51	0.64	0.64
Charlie	0.76	0.65	0.43	0.63	0.82	0.81
Delta	0.06	0.84	0.00**	0.53	0.04	0.84
Echo	0.63	0.78	0.63	0.17	0.66	0.70
Foxtrot	0.73	ϕ	0.27	ϕ	0.63	ϕ
Golf	0.50	0.65	0.50	0.41	0.52	0.66
Hotel	0.87	0.95	0.45	0.51	0.79	0.94
India	0.60	0.61	0.45	0.33	0.61	0.54
Juliet	0.79	0.83	0.63	0.53	0.80	0.79
Kilo	0.78	0.81	0.19	0.03	0.71	0.71
Lima	0.62	0.79	0.44	0.63	0.64	0.80
Mike	0.51	0.89	0.28	0.67	0.53	0.86
November	0.17	0.70	0.00**	0.52	0.07	0.70

Note: PR (Physician Rater); NP (Non-Physician)

ϕ Data missing for one non-physician rater

** Variance component equalled zero for participants

Table 26

The variance, percent variance and EP^2 for the physician raters

Design (Score)	Facets	Variance	% Variance	Ep^2
Checklist score	Candidates(p)	10.7	9.2%	0.70
	Cases (c)	12.3	10.5%	
	Raters (r)	3.2	2.7%	
	p x c	41.6	35.6%	
	p x r	0.05	0%	
	c x r	11.3	9.7%	
	p x c x r	37.7	32.3%	
Global score	Candidates(p)	0.06	5.4%	0.61
	Cases (c)	0.04	3.6%	
	Raters (r)	0.11	9.8%	
	p x c	0.28	25.0%	
	p x r	0.00	0%	
	c x r	0.16	14.3%	
	p x c x r	0.47	42.0%	
Total score	Candidates(p)	12.5	9.1%	0.71
	Cases (c)	12.9	9.4%	
	Raters (r)	7.3	5.3%	
	p x c	45.4	33.0%	
	p x r	0.00	0%	
	c x r	16.8	12.2%	
	p x c x r	42.5	30.9%	

Non-physician raters. Large variance components in the case facet was observed, which indicates there are differences in case difficulty. The variance accounted for by raters, across the three assessment formats, hovers around one percent, which demonstrates there is little difference between the two non-physician raters. Some variance is noted in the participant x rater interaction, particularly on the global rating format, which indicates some differences in how the non-physicians are evaluating the participants. Variance is observed in the case x rater interaction, principally on the global score, which indicates there is a difference between how the non-physician raters are evaluating the cases. The variance component accounted for by the p x c x r, residual interaction likely indicates that more candidates should have been selected. The variance components for the non-physician raters are located in Table 27.

Communication skills checklist.

The physician and non-physician raters evaluated the communication skills of the candidates using a 13-item checklist. A copy of the communication checklist is located in Appendix Q. The overall internal consistency, on the cases, for the physician raters equalled 0.88, 0.94, and 0.67 for the physician, non-physician, and SP raters respectively.

The internal consistency for the physician raters by case ranged between 0.56 and 0.96, while the internal consistency for the non-physician raters ranged between 0.63 and 0.94. The internal consistency for the SP raters, by case, ranged between 0.51 and 0.94. A comparison of the mean score, standard deviation and Cronbach's alpha, across the fourteen cases, between the physician, non-physician, and SP raters is located in Table 28.

Table 27

The variance, percent variance and EP^2 for the non-physician raters

Design (Score)	Facets	Variance	% Variance	Ep^2
Checklist score	Candidates(p)	11.1	8.4%	0.66
	Cases (c)	41.7	31.6%	
	Raters (r)	1.9	1.4%	
	p x c	58.4	44.2%	
	p x r	1.1	0.8%	
	c x r	1.3	0.9%	
	p x c x r, res	16.6	12.6%	
Global score	Candidates(p)	0.07	6.7%	0.57
	Cases (c)	0.26	24.8%	
	Raters (r)	0.00	0.0%	
	p x c	0.27	25.7%	
	p x r	0.05	4.8%	
	c x r	0.10	9.5%	
	p x c x r, res	0.30	28.6%	
Total score	Candidates (p)	12.9	8.8%	0.67
	Cases (c)	49.7	33.8%	
	Raters (r)	1.5	1.0%	
	p x c	61.5	41.8%	
	p x r	1.9	1.3%	
	c x r	2.1	1.5%	
	p x c x r, res	17.5	11.8%	

Table 28

A comparison between the mean(SD) and Cronbach's alpha for the physician, non-physician, and SP raters on communication checklist for the fourteen cases

Case	Mean (SD)			Cronbach's alpha		
	PR	NP	SP	PR	NP	SP
Alpha	3.5 (0.77)	3.2 (0.73)	3.8 (0.64)	0.91	0.93	0.92
Bravo	3.8 (0.59)	3.5 (0.50)	4.3 (0.50)	0.87	0.82	0.82
Charlie	3.8 (0.35)	3.3 (0.76)	4.3 (0.54)	0.82	0.93	0.94
Delta	4.0 (0.62)	3.8 (0.33)	4.1 (0.46)	0.96	0.69	0.88
Echo	3.8 (0.38)	3.9 (0.26)	3.7 (0.50)	0.88	0.63	0.90
Foxtrot	3.8 (0.34)	3.8 (0.60)*	3.8 (0.49)	0.78	0.94*	0.93
Golf	3.5 (0.37)	3.4 (0.45)	3.5 (0.24)	0.77	0.75	0.70
Hotel	3.9 (0.47)	3.7 (0.35)	4.0 (0.72)	0.79	0.68	0.91
India	3.7 (0.52)	3.5 (0.40)	3.9 (0.86)	0.85	0.77	0.96
Juliet	3.6 (0.58)	3.6 (0.48)	3.2 (0.56)	0.89	0.80	0.85
Kilo	3.9 (0.25)	4.0 (0.48)	4.0 (0.56)	0.64	0.87	0.91
Lima	3.5 (0.41)	3.3 (0.80)	3.2 (0.94)	0.82	0.93	0.94
Mike	4.0 (0.28)	4.1 (0.27)	4.5 (0.50)	0.84	0.71	0.91
November	3.7 (0.22)	3.9 (0.43)	4.1 (0.46)	0.56	0.80	0.91

Note: PR = Physician Raters, NP = Non-physician raters; SP = Standardized Patient

* Data missing for one non-physician rater

The internal consistency for the thirteen communication items is generally lower for the physician, non-physician and SP raters. The internal consistency for the physician raters, by case, ranged between -0.71 and 0.72, the internal consistency for the non-physician raters ranged between 0.22 and 0.92. The internal consistency for the SP raters, by case, ranged between -0.98 and 0.26. A comparison of the mean score, standard deviation and Cronbach's alpha, across the thirteen checklist items, between the physician, non-physician, and SP raters is located in Table 29.

The mean score by cases, across the raters, is above 3.0 on a five point scale, with several above 4.0. The mean score by checklist items, across the raters, is typically above 3.0 with several above 4.0. The only checklist item mean below 3.0 is item nine (the doctor gave the patient the opportunity to express his/her feelings or ideas in planning treatment, tests, or follow-up).

Summary Study Two

Summary statistics comparing the WAAIP physician (WP), videotape physician (VP), and two non-physician (NP) raters revealed statistically significant differences on three cases. A comparison scores between the physician raters (WP and VP) and non-physician raters showed statistically significant differences on the checklist score in four cases, differences in the global scores on three cases, and differences in the total score on three cases. The intra-class coefficient (inter-rater reliability) range low to high on the physician rater data and moderate to high on the non-physician rater data. These results could indicate the influence of pooling the WP data with the VP scores and the stability of scores with the NP raters due to their standardized training by the VP.

Table 29

A comparison between the mean(SD) and Cronbach's alpha for the physician, non-physician, and SP raters on the thirteen communication checklist items

Item	Mean(SD)			Cronbach's alpha		
	PE	NP	SP	PE	NP	SP
Understand patient	4.0 (0.33)	3.9 (0.22)	4.0 (0.17)	0.72	0.68	-0.42
Sensed feelings	3.9 (0.25)	3.9 (0.32)	3.9 (0.22)	0.53	0.75	0.14
Doctor took no notice *	4.0 (0.19)	4.2 (0.52)	3.9 (0.20)	0.02	0.84	-0.26
Doctor's response fixed *	3.9 (0.20)	4.3 (0.48)	4.0 (0.19)	0.02	0.84	0.09
Respect and courtesy	4.3 (0.09)	4.1 (0.23)	4.4 (0.11)	-0.71	0.69	-0.69
Explained problem	3.8 (0.27)	3.9 (0.28)	4.0 (0.16)	0.59	0.71	-0.68
What is wrong with Pt.	3.7 (0.28)	3.4 (0.41)	3.7 (0.29)	0.38	0.42	0.11
Explained treatment	3.6 (0.29)	3.2 (0.48)	3.6 (0.31)	0.33	0.59	0.26
Input on treatment	3.2 (0.28)	2.6 (0.28)	3.3 (0.28)	0.14	0.22	0.08
Chance to ask questions	3.4 (0.33)	3.0 (0.50)	3.6 (0.24)	0.39	0.70	-0.29
Non-technical language	3.9 (0.21)	3.9 (0.19)	4.0 (0.25)	0.31	0.26	0.14
Careful and thorough	3.7 (0.31)	3.1 (0.52)	4.0 (0.15)	0.52	0.77	-0.76
Satisfied with care	3.4 (0.27)	3.4 (0.44)	3.9 (0.17)	0.41	0.92	-0.98

Note: PR = Physician Rater, NP = Non-physician rater; SP = Standardized Patient

* Questions are reverse coded

Generalizability analyses demonstrated differences between the physician and non-physician raters. There is a higher percent variance accounted in the case facet for the non-physician raters in comparison to the physician raters suggesting that there are differences in how physicians versus non-physicians view the difficulty of the cases. Higher variance is accounted in the rater facet (especially on the global score) for the physician raters suggesting there are differences in the physician raters' scores. Less variance is accounted in the raters' facet for the non-physician raters, which is likely the result of standardized training.

The mean scores for the case and checklist items typically range between 3.0 and 4.0 on a five point scale, suggesting that the communication skills of the candidates are good. Internal consistency of the communication skills checklist is higher for each case in comparison to the individual checklist items.

CHAPTER FIVE

Discussion

The main findings of study one are: 1) the reliability of the assessment formats range from low to moderate, which calls into question the reliability of the OSCE, and 2) physician raters, standardized patients, and physician rater/SP combination accounted for approximately 20% of the variance in the assessment of clinical competency, as calculated by generalizability analyses.

The main findings of study two are: 1) there are statistically significant differences between physician and non-physician raters' scores on five cases (assessment format dependent), 2) the inter-rater reliability ranges from low to high on the physician rater data and moderate to high on the non-physician rater data, and 3) generalizability analyses revealed that the percent variance between the physician and non-physician raters varies on several of the main and interaction facets.

The primary purpose of the present study was to identify any sources of error affecting the reliability of the WAAIP OSCE that was designed to evaluate IMGs. Generalizability analyses and other statistical procedures were utilized to study the psychometric properties of the performance-based examination that was developed to evaluate clinical competency in history taking, application of physical examination skills, and the ability to communicate and/or counsel patients. A secondary purpose was to evaluate the reliability of physician versus non-physician raters to score performance-based assessments from videotape recordings of OSCE cases.

Research Question One: What are the sources of variance in clinical performance evaluation, using an Objective Structured Clinical Examination, as determined by generalizability analyses?

Four sources of variance will be discussed. These included SPs, physician raters, SP/physician rater combinations, and variance associated with assessment formats.

Sources of variance (SPs). The mean percent variance (across all three assessment formats) for SPs nested into case accounted for one fifth (approximately 20%) of the total variance (which equals 100%). This non-zero variance component for nested SPs indicates that there are differences in the scores due to the SPs selected to portray the cases. A limitation of generalizability analysis is that the specific location of where the variance is occurring can not be identified; however, observations made during study two provide evidence that differences in training protocols between the SPs trained for a case are responsible for the non-zero variance. One example will be presented to illustrate differences in training protocols.

In the Alpha case (fever) one SP sprayed his face and gown with water prior to the candidate entering the examination room. Each candidate found this SP sweaty, lying on the examination table, and looking genuinely unwell. The candidates interacting with the other SP entered the examination room to find the patient sitting in a chair, not sweaty, and not appearing unwell. When asked to move to the examination table, this SP walked effortlessly across the room and pushed himself up onto the examination table with the leg that is supposedly inflamed with thrombophlebitis (the source of the fever).

Colliver and Williams (1993) declared that the use of multiple SPs to portray a case only adds a small amount of variance to the case score. In a subsequent review of

SP-based assessments, Williams (2004) stated that research performed after the Colliver and Williams article confirmed that multiple SPs only adds a small amount of variance to the assessment. The results of the current study suggest that the protocols that were undertaken to train the two teams of SPs be carefully reviewed. [5]

Sources of variance (physician raters). The mean percent variance for the raters assigned to a case was approximately seven percent, which indicates that there are differences in candidates' scores due to the raters assigned to evaluate the case. A second rater variance component arose from the generalizability analysis (case crossed with raters assigned to case). These results are associated with the ten raters that evaluated one case during the morning session and a different case in the afternoon, indicating that the training protocols (e.g., preparation time to review case and checklist) were deficient for the new case.

Boulet et al. (2003), van der Vleuten and Swanson (1990) and Wass et al. (2001) have all advocated standardized training of examiners. Mavis and Henry (2002) declared that extensive examiner training is required if the evaluation is used for summative or high-stakes assessment. The two non-zero variance components calculated suggest that physician rater training protocols be reviewed prior to the next high-stakes examination. [6]

Sources of variance (Physician rater/SP combinations). The researcher in the current study has not found any evidence in the literature to indicate that the influence of Rater/SP combinations on candidates' scores in performance-based assessments has been

[5] No information was provided to the researcher regarding how the SPs were selected, the previous experience of the SPs in high-stakes examinations, how many SP trainers were involved, whether a physician was involved in the training process (as recommended by Barrows), or the duration of the training program (e.g., the number of hours of training depending upon the complexity of the case) was provided.

[6] No information regarding how the physician raters were selected, their experience in evaluating high-stakes performance-based assessment, or the duration of training for each case was provided.

evaluated using generalizability analysis. Observations by the videotape raters while cueing candidate/SP interactions for scoring revealed instances of the physician raters discussing the case and its presentation with the SP (most notably the Echo; risk assessment case). The mean percent variance for the SP/R combinations accounted for almost one fifth (approximately 20%) of the total variance. These results suggest that there are scoring differences between the SP/R combinations and these differences have influenced the candidates' scores.

The scoring differences might be reflective of SP case portrayal (well portrayed versus poorly portrayed). The Alpha case (fever) featured one supine, sweaty, sick SP in comparison to one sitting, non-sweaty, and not particularly ill looking SP. Statistically significant differences on the global and total scores between the two SPs were calculated. Four of the twenty-six candidates (15.4%) that interacted with the 'sweaty SP' failed, whereas five of thirteen candidates (38.5%) that interacted with the 'sitting SP' failed, furthermore, five of the candidates that did pass the case (based on attaining enough checklist items) were awarded a borderline fail score on the global assessment by the physician rater. The poor (or inaccurate) portrayal by the sitting SP may have confused or distracted some of the candidates as a patient unwell, complaining of a fever, and in the currently undiagnosed throes of thrombophlebitis is unlikely to readily cross the room and hoist one's self up onto an examination table using the effected leg.

Sources of variance (Assessment format). There are two notable observations regarding the percent variance for SPs, physician raters, and SP/R combinations. First, the percent variance is always the highest for the global rating in comparison to the checklist and total score variance. Second, the percent variance for the total score is always higher than for the checklist variance, which demonstrates the influence of the global score on the total score.

The global score was based on a 1 (poor) to five (excellent) rating system, however, each rating value was not anchored to a description of what constituted, for example in the headache case, a '3' (borderline pass). Not specifying the questions, physical assessment protocols, and/or communication skills a candidate must demonstrate to warrant a particular global rating is one potential reason for the variance associated with the global score (and its subsequent impact on the total score). Depending on the physician raters to apply their own judgement as to what represents, in the assessment of overall performance, a 'borderline fail' versus a 'borderline pass' for a particular case might be considered no different than the subjective assessments of performance that introduced variability in the traditional clinical examination format; variability the OSCE format of assessment was designed to overcome.

van der Vleuten and Swanson (1990) stated that a major concern of performance-based assessments is whether the scoring protocol or protocols developed are able to reliably capture candidate performance and translate that performance into a meaningful score. The results of the current study suggest that the global rating format be carefully evaluated. An 'overall' score for a case may be suitable if the rating points are anchored (using a consensus approach advocated by Gorter and colleagues). Or, the decision could be made to utilize multiple anchored, case-specific global scores (especially when the assessment involves more advanced candidates).

Research Question Two: Is there a difference between physician and non-physician raters in the assessment of clinical competence

Based on the results of study two, there are differences between physician and non-physician raters in the assessment of clinical competence. Caution should be taken when interpreting the physician versus non-physician results, however. The physician rater data is a combination of scores collected from the thirty-one WAAIP physicians and the scores secured from the one videotape physician rater.

A review of the study one ANOVA analyses revealed significant statistical differences between the three assigned WAAIP physician raters on eight cases (Alpha, Delta, Foxtrot, Hotel, India, Kilo, Lima, and November). In four of these cases (Delta, India, Lima, and November), the differences are found on two or more of the assessment formats. In comparison, a review of the study two ANOVA results revealed significant differences between the videotape physician rater and the two non-physician videotape raters on four cases: India (global score), Kilo (global score), November (global score), and Bravo (checklist score).

Physician rater (WP and VP) and non-physician rater (NP1 and NP2) differences were observed on the global assessment format in three cases (Delta, Golf, and India), and on the checklist and total score formats on four cases (Bravo, Delta, Echo, and Golf). Two cases, Delta and Golf, presented statistically significant differences across all three assessment formats (checklist, global, and total scores), while the Bravo case had statistically significant differences on two of the assessment formats (checklist and total scores). Differences were also noted on the checklist format in the Echo case and the global format in the India case. It could be suggested that these results establish the influence of the WAAIP physician raters' scores on the analysis between the physician and non-physician raters. The videotape physician rater was responsible for the training the two non-physician raters. The results, while not perfect, could lend support to the suggestion that the training provided by the videotape physician rater aided in the consistent scoring between the three videotape raters and the plethora of differences found in the WAAIP data reflects a corresponding lack of training of the physician raters used for the high-stakes assessment of IMGs.

The intra-class coefficient (ICC) calculations used to evaluate inter-rater reliability revealed higher coefficients and a narrower range of the coefficients for the non-physician raters results on the checklist and total scores. This is representative of the standardized training that was provided to the non-physician raters. The inter-rater reliability coefficients for the global scores are considerable lower for both the physician and non-physician scores. The results for the physician raters might be a reflection of the different expectations of performance by the physician raters. Petrusa (2004) stated that attending physicians that supervise medical students or interns might have considerably different ideas on what constitutes an acceptable performance in the assessment and management of a patient (a real patient). In the current OSCE, differing expectations of appropriate case management between the WAAIP physician raters are manifested in the global scores.

The lower inter-rater reliability coefficients (global scores) observed in the non-physician rater data might be a reflection of the non-physician raters' limitations in evaluating whether the candidate demonstrated a logical and organized approach in the evaluation of the patient. Global scales are often used to evaluate the more difficult to define skills (e.g., the fluency of the candidate's physical assessment skills) or overall performance on a particular OSCE station (Norman et al., 1991; Streiner, 1985) and extensive training of examiners is required to ensure the consistency of scoring when using a global rating system (Wass et al., 2001). Medical inexperience by the non-physician raters resulted in the low inter-rater reliability on the global scales.

Humphrey-Murto and colleagues (2005) produced similar findings when they assessed the degree of agreement between trained non-physician raters and physician raters on three of twelve OSCE stations that comprise the Medical Council of Canada Qualifying Examination Part II. Their results indicated good agreement between the physician and non-physician raters on the checklist scores but poor agreement on the global assessments.

Generalizability analyses performed on the three assessment formats revealed some curious differences between the physician and non-physician raters on the case, rater, and case by rater facets. The variance accounted in the case facet indicates a difference in case difficulty, which is typically expected. The percent variance for the checklist and total scores, however, hovered around 10% for the physician raters and over 30% for the non-physician raters. Could the physician raters, on the basis of their medical experience, view the fourteen cases as possessing approximately the same 'degree of difficulty' within the grand scheme of potential cases? It is possible that the non-physician raters lacked the subtlety to evaluate case difficulty and simply rated them as either difficult (e.g., cases that require both the asking of questions, listening to the patient's response, and the subsequent application of physical assessment and diagnostic skills) or not as difficult (e.g., counselling cases where there is little activity beyond asking questions and listening to the patient's response); hence the larger percent variance on the case facet in comparison to the physician raters.

The variance components for the rater facet are higher for the physician raters in comparison to the non-physician raters. The percent variance in the physician global score (9.8%) is considerably higher than the checklist score (2.7%) and demonstrates the influence of the global score on the total score percent variance (5.3%). Lower percent variance on the checklist variance is likely a reflection of the assessment format, which provides more structure to the evaluation, in comparison to the global score, which could be more vulnerable to the differences in physician rater subjective expectations of candidate performance. The lower percent variance results observed in the non-physician rater results could suggest the influence of the standardized training provided by the physician rater.

Finally, notable differences between the physician and non-physician raters are observed on the case x rater interaction facet. Variance indicates that there are differences in how the two categories of raters evaluated the cases. The variance components for the physician raters are higher on all three assessment formats in comparison to the non-physician raters. Physician raters might have different standards for how certain cases should be evaluated (e.g., counselling versus clinical competency) in comparison to non-physician raters who, due to their clinical inexperience as physicians have limited experience to draw upon when evaluating performance. Differences in how the non-physician raters evaluated the cases are observed in the global scores; although this might be a further demonstration of the inability of the non-physician raters to reliably use the global assessment format.

Research Question Three: Can standardized patients assess communication skills as effectively as physician raters?

The internal consistency for both the physician raters and SPs across the cases was high for both groups with the physician raters scoring slightly higher, while internal consistency scores across the checklist items were considerably lower for both groups. A series of Generalizability analyses demonstrated differences in how the physician raters and SPs were evaluating each checklist item across the fourteen cases. The results, however, should be viewed with caution due to the design of the checklist used to evaluate the communication skills of the candidates.

A generic thirteen item checklist was used to evaluate each OSCE case. The checklist utilized on a one to five rating scale (1 = strongly disagree, 2 = disagree, 3 = not sure, 4 = agree, and 5 = strongly agree). There was no 'not applicable to this case' option associated with any of the checklist items. A review of the instructions provided to the candidates prior to them entering the examination room revealed that in the majority of the cases, the candidates were simply instructed to "take a history and perform a focused physical assessment". It is unknown whether the SPs were familiar with the case-specific instructions provided to the candidates or were evaluating the candidates based on the items comprising the communication checklist. As no specific instructions were provided to the candidate to explain to the SP "what might be the matter with him/her" (item 7) or outline what "treatment, tests or other follow-up is going to happen" (item 8) the candidates might not have performed these communication tasks choosing to further evaluate the chief complaint of the case (e.g., severe flu). Could the candidates' failure to address a communication item they had not been explicitly instructed to perform result in a low score on that item because of SP expectations?

The physician raters were provided with a copy of the instructions presented to the candidates but might also have been unable to resolve how an item the candidate had not been clearly instructed to address could be scored. A recourse to score the item as a '3' (not sure) is not the same as 'not applicable to this case', however as mentioned, this option was not available. It will be proposed that the context-specificity of the case must be considered when developing a communication checklist, a recommendation that has support in the literature (Guiton et al., 2004).

The high overall internal consistency results (Cronbach's alpha) for both the SP and physician raters would appear to support the reliability of the instrument used to measure the communication skills of the candidates as would the high internal consistency coefficients calculated for the majority of the fourteen cases. Despite these results, the lower reliability coefficients for the thirteen checklist items across the fourteen cases and the percent variance components calculated with generalizability analysis subvert the assertion that the communication checklist utilized to evaluate the candidates in the current study is a reliable assessment instrument.

Research Question Four: Can the characteristics, strengths and limitations of an OSCE be identified in order to create a model for the reliable assessment of competency in medicine?

The analysis of the current OSCE indicates that considerable variance is associated with the nested SPs, the assigned physician raters, and the assigned

SP/physician rater combinations. Extensive recommendations for the preparation of SPs, training of examiners, case design and development can be readily found in the medical education literature. Could it be proposed that, based on the OSCE and SP literature, a model for the reliable assessment of performance-based clinical competency assessment has already been created and only requires those implementing an OSCE to follow the outlined recommendations?

OSCEs were designed to be a more controlled assessment of clinical competence by removing the variability introduced by patients and examiners that was found in the traditional clinical examination. In the current OSCE, the skill level of the candidates might have influenced reliability due to the choice of assessment formats used to assess clinical competency. The participants were medical doctors; many with extensive post-graduate training and several possessing many years of experience working as a family physicians in his/her respective country of origin. It should also be take into consideration that the WAAIP project was developed to evaluate whether the candidates knowledge and clinical skills were sufficiently competent to permit them to bypass a residency program and be integrated into the physician workforce.

Clinical competency was evaluated using a case-specific checklists with one overall global rating. While this could be considered an appropriate assessment protocol for medical students in the early years of training (Norman et al., 1991; van der Vleuten & Swanson, 1990), it should be questioned whether the assessment format utilized was the most appropriate assessment format for the candidates in the current study. Global ratings have been recommended for more advanced participants in performance-based evaluations (Norman et al., 1991). While the reliability of global rating format used in the current OSCE has been questioned due to the lack of anchoring of rating points, case-specific global ratings (with anchor points) used by extensively trained raters might provide a more rounded evaluation of clinical competence.

The strength of the checklist format is that a well-designed checklist can provide structure for the assessment while identifying the specific guidelines required for the appropriate management of the case. One consideration regarding the checklist format, in the current OSCE, must be focused on the equal weighting of the case-specific items. Each item was scored on a 'yes' (the candidate addressed the item) or 'no' (the candidate did not address the item) basis. In the Bravo (headache) case, eliciting that the patient's parents were divorced when she was 15 is worth the same point value as performing a critical physical assessment test (Kernig's sign) that is used to evaluate for meningial irritation. One strategy might involve the incorporation of the 'key actions' or 'key features' format of assessment (Page & Bordage, 1995) and then identifying which candidates are addressing the critical components of the case.

There were several examples of candidates failing a case despite having attained a borderline pass on the global rating (or candidates passing a case despite a fail or borderline fail rating). There were also examples of candidates having failed a case despite making the correct diagnosis (while others passed the case despite an incorrect diagnosis). Is it possible that some of the candidates approached the OSCE as though they would approach a real patient in the clinical setting by only asking the critical questions and performing the critical physical assessments based on the chief complaint and the SPs responses to specific questions/assessments. For this reason, they were unable to collect the required number of checklist items to pass the case's minimum performance level? Could it also be that some of the examiners rated the performance based on the number of items addressed, regardless of whether the relevant items were addressed?

The purpose of assessment is to provide inferences about the ability or competency of the candidates; inferences that extend beyond the sample of cases or cases included (van der Vleuten & Swanson, 1990). Petrusa (2002) recommended that future research on the most appropriate assessment format for the skill level of the performer is essential. Petrusa (2002) further recommended that future research in performance-based assessment focus on data collection and scoring in order to ensure "that better performance has a better score" (p. 705).

Research Question Five: Are there consistencies in the types of errors of measurement that are introduced by raters, SPs, and case variability?

Generalizability analyses indicated that there are differences in candidates' scores due to SPs, raters, and SP/rater combinations. The limitation of Generalizability is that the variance only provides evidence that SPs, raters, and SP/rater combinations have influences scores – but can not identify the location. Observations made during the physician/non-physician raters study provides some clues.

SPs. Inconsistency between the two SPs' portrayals was observed in both the Delta case (informed consent) and Hotel (flu symptoms) case, while inaccuracy in SP performance was observed in the Bravo case (headache). Differences in the portrayal of illness were also noted in the Lima case (vomiting). Based on the observed irregularities in SP presentation, it is suggested that a quality assurance assessment, along the lines of

Tamblyn (1989) be undertaken. Tamblyn evaluated SP accuracy within the framework of an end of the year OSCE involving final year medical students. While it was concluded that there was no relationship between SP accuracy score and overall candidate competence score, there was one case where SP accuracy was high (98%) but an error in the presentation of one critical item had consequences for diagnosis and case management (Tamblyn, 1989).

Research utilizing the Tamblyn framework could be incorporated as a quality assurance protocol; potentially as an extension from the quality assurance methods proposed by Boulet and colleagues (2003), although the time and cost required to undertake as extensive an assessment would be prohibitive. One potential strategy for overcoming time and cost constraints could be to randomly select candidate/SP interactions from cases where the fail rate is unusually high (e.g., the Charlie case; 49% fail rate), when a large number of candidates interacting with a particular SP are not correctly diagnosing the case (e.g., the Bravo case – SP not portraying photosensitivity as instructed in the SP manual), when a large number of candidates are correctly diagnosing the illness or injury but are still failing the case (e.g., the Juliet case where extensive the patient history and physical assessment components might not be suitable to the assessment time-frame), or when documentation of OSCE observers are noting differences in SP portrayal (e.g., the Alpha case – the 'prone and sweaty SP' versus the 'sitting in a chair and not sweaty SP').

Case Design. From a case design standpoint, each case (with the exception of November) was comprised of a candidate/SP interaction followed by a post-encounter probe. The raters on the Delta (informed consent) appeared confused regarding the exact purpose of the post-encounter-probe. The Hotel case had a three question post-encounter probe that accounted for 20 possible points, which proved to be difficult to complete in the two minutes provided. Seven of the cases had 35 or more checklist items to be addressed within a ten minute time frame. Furthermore, it is unknown whether these cases were designed for or pilot-tested with a ten-minute timeframe. It is also unknown whether these case were specifically designed for high-stakes examination purposes or whether they were primarily designed for formative assessment or for training to familiarize foreign trained physicians with OSCEs as many are unfamiliar with this examination format (Andrew & Bates, 2000).

Raters. Based on observations by the videotape raters, it appeared that some of the raters may have been inadequately prepared to evaluate this high-stakes performance-based examination. Notable examples of poor preparation occurred in the two-minute post-encounter probe. Raters in the Hotel case allowed the candidates to explain their differential diagnoses even though no points were given for explaining why a diagnosis was being proposed. Raters in the India case asked the candidates whether they wanted the x-ray or blood test results. If the candidate asked for the x-ray results, approximately one minute and 45 seconds was taken by two of the raters to read the x-ray results (for which no points were included on the checklist) leaving approximately fifteen seconds for the candidate to peruse the laboratory results. In the Juliet case, one examiner was observed to ask the candidates if they wanted to add "anything else" to their management recommendations for the hand problem patient. The most egregious performance with the PEP was in the Mike case (urinary problem). One examiner completely misunderstood the "one-minute left" warning during the ten minute candidate/SP interaction, and leapt (literally) into the video frame holding the plastic prostate simulator for the candidate to palpate and list the findings (PEP question 1). In addition, the simulator proved to be adept at changing from the correct setting to adjacent settings, so time was spent by the rater resetting the simulator and holding the setting in place for the candidates to palpate.

van der Vleuten (1996) cautioned against using performance-based evaluations that did not include physician judgment, as physician raters are considered to be more familiar with the logical sequencing of the history taking and physical examination processes. Physician raters are also considered betters at evaluating the technical proficiency of physical assessment procedures (van der Vleuen & Swanson, 1990). Despite these perceived advantages, examiner inconsistency is the largest threat to the reliability of an assessment (Downing, 2004). For this reason, standardized training of the physician raters on the use the assessment instruments is critical (Boulet et al., 2003; van der Vleuten & Swanson, 1990; Wass et al., 2001). In the case of the current study, it might be suggested that while physician expertise is valuable the advantage is neutralized by poor preparation.

Reliability and Sources of Error

The primary purpose of the current study was to identify the sources of error affecting the reliability of an OSCE developed to evaluate the clinical competency of foreign trained physicians. The results call into question the reliability of the OSCE. The generalizability coefficient (Ep^2) for the checklist and total scores based on the two-facet designs (SP nested into case, raters assigned to cases, and rater/SP combination) all have a moderately high correlation (ranging from 0.82 to 0.84). The Ep^2 for the global scores are lower, ranging from 0.71 to 0.74 for the two-facet designs. These generalizability coefficients may appear impressive, however should be interpreted with caution in consideration of the results from the variance components analyses. Furthermore, Cronbach et al. (1972) recommended the reporting of variance components rather than presenting generalizability coefficients. The non-zero variance components approaching 20% for each facet, which indicates that SPs, raters, and rater/SP combinations are influencing candidates' scores.

Delimitations of the current research

There are several delimitations to the current research study. The population was delimited to foreign trained physicians (the cases were not tested by physicians, residents, or medical students trained in Canada). Furthermore, the candidates who participated in the WAAIP project are but a small sample of the foreign trained physicians registered in Western Canada. Another cohort of foreign trained physicians might have performed better (or worse) on the same OSCE.

The SPs are representative one training facility (Medical Skills Centre, University of Calgary). No information was provided on how this training facility selected or trained the SPs. Likewise, the physician raters selected and trained are also a small representative of the physician raters used to evaluate high-stakes OSCEs. No information was provided on how these physician raters were selected and trained or their previous experience as raters in high-stakes OSCEs. Finally, the fourteen cases selected are also representative of the population of all possible cases. No information was provided to indicate whether they were pilot-tested with a representative sample of participants prior to their use in the WAAIP OSCE.

Limitations of the current research

There are several limitations in the current research. The sample size, for both studies, is not optimal for running a complex statistical procedure such as generalizability analysis. One rule of thumb suggests that the ideal ratio between participants and variables is ten to one; meaning 140 candidates should have been tested on the current OSCE. Only thirty-nine candidates were selected for the WAAIP program, resulting in 540 scores for each assessment format. While only fifteen candidates were selected for study two, each candidate's performance was scored by three raters, which provided 630 scores for each assessment format.

Another limitation, from a statistical standpoint, is the unequal participant size for each track (n = 11, 13, and 15, respectively) in addition to the unequal number of participant interactions with the SPs (n = 13 versus n = 26). Any outlier scores will be better absorbed in the larger sample sizes while outlier scores could unduly influence mean scores in the smaller sample sizes. From a rater standpoint, only having one rater per station did not allow for the assessment of inter-rater reliability in the WAAIP OSCE. Furthermore, not nesting the raters was a limitation as ten of the physician raters were assigned to different cases with not enough time to prepare for the new case.

Recommendations future research

Future research should compare the performance of foreign trained physicians with physicians, residents, and medical students trained in Canada or the United States on the same OSCE cases to ascertain whether there are differences in performance and, if so, where (e.g., physical assessment or diagnostic skills).

Assessment of the validity (e.g., predictive or concurrent) of an OSCE should also be addressed in future research. Friedman Ben-David (2003) stated that further work is required to clarify what is meant by validity in regards to performance assessment. According to Downing (2003), validity is approached as a hypothesis and looks to multiple sources of empirical evidence to support or refute the hypothesis. Swanson (1987) and van der Vlueten and Swanson (1990) stated that the purpose of an assessment is to provide inferences about the ability or competency of the candidates – inferences that extend beyond the sample of cases or cases included in the examination. Further research should assess candidates in the operational environment (e.g., a subsequent clinical rotation) and evaluate whether the OSCE results could predict the degree of success. A final recommendation for future research revolves around the use of non-physician raters and establishing whether they can be trained to reliably use global scales to assess physician performance.

Conclusions

The amount of variance associated with SPs, physician raters, and their combinations in the current OSCE illustrates the typical errors of measurement in performance-based assessments. Observations made during the physician/non-physician rater study revealed several problems with SPs, raters, case design and general examination administration. The evaluation of clinical competency using performance-based assessment requires an attention to detail across all phases of examination design and implementation; not just the "front end" of identifying competencies and outlining candidate selection criteria.

The following criteria should be employed in OSCE development. A carefully developed table of specifications will ensure that content validity is enhanced. The skills and competencies that will be the focus of evaluation must be reflected in that table of specifications. A meticulous task-analysis which identifies the critical skills of the task and the essential relationships between these critical skills follows. The development of the case scenario that encompasses the skills and competencies to be evaluated is essential. The case should be based on common chief complaints and be moderately difficult as cases that are too easy or too hard are poor discriminators of competency.

The scenario must be designed to elicit critical actions and responses to clinical problems. If the purpose of the case is to evaluate whether the candidate can manage a particular cardiac arrhythmia using a specific algorithm, the relevant features must be embedded into the scenario. Addressing the logical flow of the scenario and the manifestation of signs and symptoms is mandatory, so that the candidate can demonstrate clinical competency within the scope of the case.

Careful consideration of the skill level of the candidates and the expected standard of performance for that skill level is essential. The assessment of candidate skill level and the expected standard of performance underlie clinical case development and help to inform the selection of the most suitable assessment protocol(s) for the skill level of candidate being evaluated. Cutoff scores for pass/fail should be set based on standard minimum performance level (MPL) procedures such as the Ebel method.

Regardless of which assessment format is selected (checklist, global rating, key component, or a combination), a consensus approach using a team of physician educators should be utilized. An instructional designer who is well-versed in educational assessment should facilitate the entire examination development process with the content specialists. Furthermore, an instructional designer with experience in developing simulations for training and evaluation may be employed to guide the development of the cases in conjunction with subject matter experts and interact with these medical specialists to ensure that the case and SP portrayal of that case is accurate.

The importance of training SPs and physician raters has been repeatedly raised throughout this dissertation. For example, many resources can be accessed for SP training (e.g., Barrows, 1999). Pilot-testing the cases with a group of candidates that are representative of the population to be assessed should be done to ensure the cases flow in a logical manner and that the tasks and critical actions/responses can be performed within the designated time frame for the case.

Upon completion of the examination (or during if problems arise), feedback from the SPs and physician raters on how the cases ran and what unforeseen problems were encountered must be elicited. After collecting the performance data an explicit psychometric evaluation must be undertaken (the statistical procedures used in the current research are recommended) as part of a quality assurance assessment so that subsequent performance-based assessments can be designed more deliberately to increase reliability and reduce the sources of error. Such detailed procedures will thus enhance the reliability of the OSCE format and accordingly reduce errors of measurement.

References

Adamo, G. (2003). Simulated and standardized patients in OSCEs: Achievements and challenges 1992-2003. *Medical Teacher, 25*(3), 262-270.

Andrew, R., & Bates, J. (2000). Program for licensure for international medical graduates in British Columbia: 7 years' experience. *Canadian Medical Association Journal, 162*(6), 801-803.

Barrows, H. (1971). *Simulated patients (programmed patients): The development and use of a new technique in medical education.* Springfield, Illinois: Charles C. Thomas.

Barrows, H. (1987). *Simulated (Standarized) Patients and Other Human Simulations.* Chapel Hill, North Carolina: Health Sciences Consortium.

Barrows, H. (1993). An overview of the uses of standardized patients for teaching and evaluating clinical skills. *Academic Medicine, 68*(6), 443-451.

Barrows, H. (1999). *Training standardized patients to have physical findings.* Springfield, Illinois: Southern Illinois University School of Medicine.

Boulet, J., McKinley, D., Norcini, J., & Whelan, G. (2002). Assessing the comparability of standardized patient and physician evaluations of clinical skills. *Advances in Health Sciences Education, 7*, 85-97.

Boulet, J., McKinley, D., Whelan, G., & Hambleton, R. (2003). Quality assurance methods for performance-based assessments. *Advances in Health Sciences Education, 8*(1), 27-47.

Brennan, R. (1994). Variance components in generalizability theory. In C. Reynolds (Ed.), *Cogntive Assessment: A Multidisciplinary perspective.* New York: Plenum Press.

Burns, K. (1998). Beyond classical reliability: Using generalizability theory to assess dependability. *Research in Nursing and Health, 21,* 83-90.

Canadian Resident Matching Service. (2008). Main Residency Match (R1). Eligibility Provincial Restrictions. Retrieved http://www.carms.ca/eng/r1_eligibility_prov_e.shtml . Accessed March 1, 2008

Carraccio, C., & Englander, R. (2000). The Objective Structured Clinical Examination: A step in the direction of competency-based evaluation. *Archives of Pediatric and Adolescent Medicine, 154,* 736-741.

Cohen, R., Rothman, A., Poldre, P., & Ross, J. (1991). Validity and generalizability of global ratings in an Objective Structured Clinical Examination. *Academic Medicine, 66*(9), 545-548.

Colliver, J., Robbs, R., & Vu, N. (1991). Effects of using two or more standardized patients to simulate the same case of case means and case failure rates. *Academic Medicine, 66*(10), 616-618.

Cronbach, L., Gleser, G., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Crossley, J., Davies, H., Humphris, G., & Jolly, B. (2002). Generalisability: a key to unlock professional assessment. *Medical Education, 36*(10), 972-978.

Crossley, J., Humphris, G., & Jolly, B. (2002). Assessing health professionals. *Medical Education, 36*(9), 800-804.

Crutcher, R., Banner, S., Szafran, O., & Watanabe, M. (2003). Characteristics of international medical graduates who applied to the CaRMS match. *Canadian Medical Association Journal, 168*(9), 1119-1123.

Cunnington, J., Neville, A., & Norman, G. (1997). The risks of thoroughness: Reliability and validity of global ratings and checklists in an OSCE. *Advances in Health Sciences Education, 1*(3), 227-233.

Dauphinee, W. (2005). Physician migration to and from Canada: the challenge of finding the ethical and political balance between the individual's right to mobility and recruitment to underserved communities. *Journal of Continuing Education in the Health Professions, 25*(1), 22-29.

Dimitrov, D. (2002). Reliability: Arguments for multiple perspective and potential problems with generalization across studies. *Educational and Psychological Measurement, 62*(5), 783-801.

Downing, S. (2003a). Item response theory: Applications of modern test theory in medical education. *Medical Education, 37*(8), 739-745.

Downing, S. (2003b). Validity: On the meaningful interpretation of assessment data. *Medical Education, 37*(9), 830-837.

Downing, S. (2004). Reliability: On the reproducibility of assessment data. *Medical Education, 38*(9), 1006-1012.

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

Ebel, R., & Frisbe, D. (1986). *Essentials of Educational Measurement. 4th ed.* Englewood Cliffs, New Jersey: Prentice Hall.

Friedman Ben-David, M. (2003). Life beyond OSCE. *Medical Teacher, 25*(3), 239-240.

Glassman, P., Luck, J., O'Gara, E., & Peabody, J. (2000). Using standardized patients to measure quality: Evidence from the literature and a prospective study. *Joint Commision Journal on Quality Improvement, 26*, 644-653.

Gorter, S., Rethans, J., Scherpbier, A., van der Heijde, D., Houben, H., van der Vleuten, C., & van der Linden, S. (2000). Developing case-specific checklists for standardized-patient-based assessments in internal medicine: A review of the literature. *Academic Medicine, 75*(11), 1130-1137.

Guiton, G., Hodgson, C., Delandshere, G., & Wilkerson, L. (2004). Communication skills in standardized patient assessment of final-year medical students: A psychometric study. *Advances in Health Sciences Education, 9*, 179-187.

Hambleton, R. (1989). Principles and selected applications of Item Response Theory. In R. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 147-200). New York: Macmillan Publishing Company.

Harden, R., & Gleeson, F. (1979). Assessment of clinical competence using an objective structured clincial examination (OSCE). *Medical Education, 13*, 41-54.

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

Hilliard, R., & Tallett, S. (1998). The use of objective structured clinical examination with postgraduate residents in pediatrics. *Archives of Pediatric and Adolescent Medicine, 152*, 179-184.

Hodges, B. (2003). Validity and the OSCE. *Medical Teacher, 25*(3), 250-254.

Humphrey-Murto, S., Smee, S., Touchie, C., Wood, T., & Blackmore, D. (2005). A comparison of physician examiners and trained assessors in a high-stakes OSCE setting. *Academic Medicine, 80*(10:Supplement), S59-S62.

Martin, I., & 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. *Medical Education, 36*(5), 418-425.

Martin, J., Reznick, R., Rothman, A., Tamblyn, R., & Regehr, G. (1996). Who should rate candidates in an Objective Structured Clinical Examination? *Academic Medicine, 71*(2), 170-175.

Mavis, B., & Henry, R. (2002). Between a rock and a hard place: finding a place for the OSCE in medical education. *Medical Education, 36*(5), 408-409.

McLaughlin, K., Gregor, L., Jones, A., & Coderre, S. (2006). Can standardized patients replace physicians as OSCE examiners? [Electronic Version]. *BMC Medical Education*, 6 from http://www.biomedcentral.com/I472-6920/6/12.

Miller, G. (1990). The assessment of clinical skills competence performance. *Academic Medicine, 65*(9 Suppl), S63-S67.

Muller, E., Harik, P., Margolis, M., Clauser, B., McKinley, D., & Boulet, J. (2003). An examination of the relationship between clinical skills examination performance and performance on USMLE Step 2. *Academic Medicine, 78*(10 Suppl), S27-29.

Murray, D., Boulet, J., Kras, J., McAllister, J., & Cox, T. (2005). A simulation-based acute skills performance assessment for anesthesia training. *Anesthesia and Analgesia, 101*, 1127-1134.

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

Norcini, J., & Mazmanian, P. (2005) Physician migration, education and health care. *Journal of Continuing Education in the Health Professions, 25*(1), 4-7.

Norman, G., van der Vleuten, C., & De Graaff, E. (1991). Pitfalls in the pursuit of objectivity: Issues of validity, efficiency and acceptability. *Medical Education, 25*, 119-126.

Page, G., & Bordage, G. (1995). The Medical Council of Canada's key features project: A more valid written examination of clinical decision-making skills. *Academic Medicine, 70*, 104-110.

Peitzman, S. (2001). Physical diagnosis findings among persons applying to work as standardized patients. *Academic Medicine, 76*, 383.

Petrusa, E. (2002). Clinical performance assessments. In G. Norman, C. van der Vleuten & D. Newble (Eds.), *International Handbook of Research in Medical Education* (Vol. 2, pp. 673-709). Dordrecht: Kluwer Academic Publishers.

Petrusa, E. (2004). Taking standardized patient-based examination to the next level. *Teaching and Learning in Medicine, 16*, 98-110.

Petrusa, E., Blackwell, T., & Ainsworth, M. (1990). Reliability and validity of an objective structured clinical examination for assessing clinical performance of residents. *Archives of Internal Medicine, 150*, 573-577.

Reznick, R., Regehr, G., Yee, G., Rothman, A., Blackmore, D., & Dauphinee, D. (1998). Process-rating forms versus task-specific checklists in an OSCE for medical licensure. Medical Council of Canada. *Academic Medicine, 73*(10 Suppl), S97-S99.

Shavelson, R., & Webb, N. (1991). *Generalizability Theory: A Primer*. Newbury Park, California: Sage Publications.

Shavelson, R., & Webb, N. (2006). Generalizability Theory. In J. Green, G. Camilli, P. Elmore, A. Skukauskaite & E. Grace (Eds.), *Handbook of Complimentary Methods in Education Research*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Shavelson, R., Webb, N., & Rowley, G. (1989). Generalizability Theory. *American Psychologist, 44*(6), 922-932.

Shavelson, R., Webb, N., & Rowley, G. (1992). Generalizability theory. In A. Kazdin & E. Alan (Eds.), *Methodological issues & strategies in clinical research.* (pp. 233-256). Washington, DC: American Psychological Association.

Shea, J., & Fortna, G. (2002). Psychometric methods. In G. Norman, C. van der Vleuten & D. Newble (Eds.), *International Handbook of Research in Medical Education* (Vol. 1, pp. 97-126). Dordrecht: Kluwer Academic Publishers.

Shumway, J., & Harden, R. (2003). AMEE Guide No. 25: The assessment of learning outcomes for the competent and reflective physician. *Medical Teacher, 25*(6), 569-584.

Sloan, D., Donnelly, M., Schwartz, R., Vasconez, H., Plymale, M., & Kenady, D. (1998). Critical assessment of the head and neck clinical skills of general surgery residents. *World Journal of Surgery, 22*(3), 229-235.

Streiner, D. (1985). Global rating scales. In V. Neufeld & G. Norman (Eds.), *Assessing Clinical Competence* (pp. 119-141). New York: Springer Publishing Company.

Streiner, D., & Norman, G. (1989). Reliability. In *Health Management Scales: A Practical Guide to their Development and Use* (pp. 79-96). New York: Oxford University Press.

Strube, M. (2000). Reliability and generalizability theory. In L. Grimm & P. Yarnold (Eds.), *Reading and Understanding Multivariate Statistics.* Washington, DC.: American Psychological Association.

Swanson, D. (1987). A measurement framework for performance-based tests. In I. Hart & R. Harden (Eds.), *Further developments in assessing clinical competence.* (pp. 13-45). Montreal: Can-Heal.

Tamblyn, R. (1989). *The use of the standardized patient in the measurement of clinical competence: The evaluation of selected measurement properties.*, McGill University, Montreal.

Tamblyn, R., Klass, D., Schnabl, G., & Kopelow, M. (1991). The accuracy of standardized patient presentation. *Medical Education, 25*(2), 100-109.

Thomas, J., Nelson, J., & Silverman, S. (2005). *Research Methods in Physical Activity* (5th ed.). Champaign, IL: Human Kinetics.

van der Vleuten, C. (1996). The assessment of professional competence: Developments, research and practical implications. *Advances in Health Sciences Education, 1*, 41-67.

van der Vleuten, C. (2000). Validity of final examinations in undergraduate medical training. *British Medical Journal, 321*(7270), 1217-1219.

van der Vleuten, C., Norman, G., & De Graaff, E. (1991). Pitfalls in the pursuit of objectivity: Issues of reliability. *Medical Education, 25*, 110-118.

van der Vleuten, C., & Swanson, D. (1990). Assessment of clinical skills with standardized patients: State of the art. *Teaching and Learning in Medicine, 2*, 58-76.

Violato, C., Marini, A., & Lee, C. (2003). A validity study of expert judgement for setting cut-off scores on high-stakes credentialing examinations using cluster analysis. *Evaluation and the Health Professions, 26*(1), 59-72.

Violato, C., Marini, A., & McDougall, E. (1998). *Assessment of Classroom Learning*. Calgary, Alberta: Detselig Enterprise, Ltd.

Violato, C., & Baig, L. (2006). Psychometric Report of the Western Alliance for Assessment of International Physicians Project. (pp. 68): University of Calgary.

Vu, N., Barrows, H., Marcy, M., Verhulst, S., Colliver, J., & Travis, T. (1992). Six years of comprehensive, clinical, performance-based assessment using standardized patients at the Southern Illinois University School of Medicine. *Academic Medicine, 67*(1), 42-50.

Wass, V., van der Vleuten, C., Shatzer, J., & Jones, R. (2001). Assessment of clinical competence. *Lancet, 357*(9260), 945-949.

Weller, J., Robinson, B., Jolly, B., Watterson, L., Joseph, M., Bajenov, S., Haughton, A., & Larsen, P. (2005). Psychometric characteristics of simulation-based assessment in anaesthesia and accuracy of self-assessed scores. *Anesthesia, 60,* 245-250.

Western Alliance for Assessment of International Physicians. (2006). Project Objective. Retrieved June 2, 2006, from http://www.waaip.ca/index.htm

Whelan, G., Boulet, J., McKinley, D., Norcini, J., van Zanten, M., Hambleton, R., Burdick, W., & Peitzman, S. (2005). Scoring standardized patient examinations: Lessons learned from the development and administration of the ECFMG Clinical Skills Assessment (CSA). *Medical Teacher, 27*(3), 200-206.

Williams, R. (2004). Have standardized patients examinations stood the test of time and experience? *Teaching and Learning in Medicine, 16*(2), 215-222.

Winckel, C., Reznick, R., Cohen, R., & Taylor, B. (1994). Reliability and construct validity of a structured technical skills assessment form. *American Journal of Surgery, 167*(4), 423-427.

Appendix A

Ethics Approval

FACULTY OF MEDICINE | UNIVERSITY OF CALGARY

OFFICE OF MEDICAL BIOETHICS
Room 93, Heritage Medical Research Bldg
3330 Hospital Drive NW
Calgary, AB, Canada T2N 4N1
Telephone: (403) 220-7990
Fax: (403) 283-8524
Email: omb@ucalgary.ca

2006-07-19

Dr. C. Violato
Department of Applied Psychology
University of Calgary
EDT 420
Calgary, Alberta

Dear Dr. C. Violato:

RE: Reliability, Validity and Sources of Errors in Assessing Physician Performance in an Objective Structured Clinical Examination: A Generalizability Theory Analysis

Ethics ID: E-20343

Student: A. Vallevand

The above-noted proposal including the Research Proposal and the Consent Form has been submitted for Board review and found to be ethically acceptable.

Please note that this approval is subject to the following conditions:
(1) appropriate procedures for consent for access to identified health information have been approved;
(2) a copy of the informed consent form must have been given to each research subject, if required for this study;
(3) a Progress Report must be submitted by July 19, 2007, containing the following information:

- i) the number of subjects recruited;
- ii) a description of any protocol modification;
- iii) any unusual and/or severe complications, adverse events or unanticipated problems involving risks to subjects or others, withdrawal of subjects from the research, or complaints about the research;
- iv) a summary of any recent literature, finding, or other relevant information, especially information about risks associated with the research;
- v) a copy of the current informed consent form;
- vi) the expected date of termination of this project.

4) a Final Report must be submitted at the termination of the project.

Please note that you have been named as the principal collaborator on this study because students are not permitted to serve as principal investigators. Please accept the Board's best wishes for success in your research.

Yours sincerely,

Glenys Godlovitch, BA(Hons), LLB, PhD
Chair, Conjoint Health Research Ethics Board

GG/sg
c.c. Dr. B. MacIntosh (information) Research Services A. Vallevand (Student)
Office of Information & Privacy Commissioner

Appendix B

Consent Form for WAAIP Candidates

Western
Alliance for
Assessment of
International
Physicians

CONSENT FOR RELEASE OF PERSONAL INFORMATION

I, ______________________________________, consent to the release of
(Name)

all documentation and assessment results to the Western Alliance for Assessment of International Physicians for the purpose of statistical analysis, program reporting and research. I understand that I will be asked to complete some questionnaires to provide information about my background. My name will be kept confidential at all times.

I understand that some of the assessments will be video taped and the results will be coded in such a way that the identity of the participants will not be revealed from the data. I understand that the results of this program may be published or reported to government agencies or scientific groups, but my name will not be linked to published results.

I acknowledge that I have been made aware of the reasons for the disclosure of the above information, and the risks and benefits associated with consenting to its release.

I understand that this consent is irrevocable.

Date: ___________________________ Valid Until: _____________________

Signature: _______________________ Print Name: ____________________

Appendix C

The Objective Structured Clinical Examination (OSCE) Cases

Patient Name and Presentation	Purpose of Case
Alpha Fever	• History Taking • Physical Assessment
Bravo Headache	• History Taking • Physical Assessment
Charlie Infection	• History Taking • Physical Assessment • Promptness of management and Information Sharing
Delta Informed Consent	• Information Sharing regarding required surgical procedure • Obtaining Informed Consent for surgery
Echo Risk assessment and disease diagnosis	• History Taking • Risk Assessment • Counseling
Foxtrot Metastasis of cancer	• History Taking • Information Sharing • Counseling (breaking bad news)
Golf Shortness of Breath	• History Taking • Physical Assessment • Information Sharing

Patient Name and Presentation	Purpose of Case
Hotel Flu Symptoms	• History Taking • Physical Assessment • Information Sharing
India Fatigue	• History Taking • Communication Skills
Juliet Problems with hand	• History Taking • Physical Assessment
Kilo Personality Changes	• History Taking • Counseling with risk management for suicidal tendencies
Lima Vomiting	• Immediate Assessment and Treatment • History Taking • Physical Assessment and emergency management
Mike Urinary tract	• History Taking • Physical Assessment • Counseling
November Cardiac counseling	• History Taking • Counseling • Risk assessment and management

Appendix D

Summary Statistics for the Checklist scores (Tracks 1, 2, and 3)

Case	Items	Track	Mean	SD	SEM	Min	Max	Range
Alpha	43	1 (n = 13)[a]	24.3	4.2	1.2	18	31	13
Fever		2 (n = 15)[b]	25.9	5.6	1.5	12	36	24
		3 (n = 11)[b]	28.6	6.3	1.9	13	37	24
		Total (n = 39)	26.1	5.6	0.9	12	37	25
Bravo	41	1 (n = 13)[a]	22.8	5.6	1.6	12	31	19
Headache		2 (n = 15)[b]	27.7	4.4	1.1	20	34	14
		3 (n = 11)[b]	25.6	6.2	1.9	16	35	19
		Total (n = 39)	25.5	5.6	0.9	12	36	24
Charlie	28	1 (n = 13)[a]	13.5	2.4	0.7	10	17	7
Infection		2 (n = 15)[b]	15.3	2.7	0.7	11	20	9
		3 (n = 11)[b]	13.0	2.7	0.8	9	17	8
		Total (n = 39)	14.1	2.7	0.4	9	20	11
Delta	13	1 (n = 13)[a]	8.7	1.2	0.3	7	10	3
Informed		2 (n = 15)[b]	6.4	1.9	0.5	4	10	6
Consent		3 (n = 11)[b]	8.7	1.2	0.4	6	10	4
		Total (n = 39)	7.8	1.9	0.3	4	10	6

Case	Items	Track	Mean	SD	SEM	Min	Max	Range
Echo	36	1 (n = 13)[a]	21.5	3.7	1.0	16	29	13
Risk		2 (n = 15)[b]	21.8	3.8	1.0	15	28	13
Assessment		3 (n = 11)[b]	22.5	2.7	0.8	19	28	9
		Total (n = 39)	21.9	3.4	0.5	15	29	14
Foxtrot	16	1 (n = 13)[a]	7.6	2.7	0.7	4	12	6
Metastasis		2 (n = 15)[b]	8.3	2.1	0.5	5	12	7
		3 (n = 11)[b]	7.7	3.0	0.9	3	13	10
		Total (n = 39)	7.9	2.5	0.4	3	13	10
Golf	33	1 (n = 13)[a]	19.0	2.6	0.7	15	23	8
Shortness of		2 (n = 15)[b]	20.8	2.4	0.6	18	25	7
Breath		3 (n = 11)[b]	21.2	3.6	1.1	17	27	10
		Total (n = 39)	20.3	2.9	0.5	15	27	12
Hotel	40	1 (n = 13)[a]	21.9	2.7	0.8	17	26	9
Flu		2 (n = 15)[b]	19.7	3.2	0.8	16	28	12
Symptoms		3 (n = 11)[b]	22.1	3.5	1.1	16	27	11
		Total (n = 39)	21.1	3.3	0.5	16	28	12

Case	Items	Track	Mean	SD	SEM	Min	Max	Range
India	32	1 (n = 13)[a]	21.7	3.7	1.0	14	25	11
Fatigue		2 (n = 15)[b]	16.2	3.4	0.9	10	23	13
		3 (n = 11)[b]	16.2	2.0	0.6	13	19	6
		Total	18.0	4.1	0.7	10	25	15
		(n = 39)						
Juliet	35	1 (n = 13)[a]	21.3	4.1	1.2	12	27	15
Hand		2 (n = 15)[b]	19.3	2.9	0.8	13	23	10
Problems		3 (n = 11)[b]	18.7	3.8	1.1	12	26	14
		Total	19.8	3.7	0.6	12	27	15
		(n = 39)						
Kilo	35	1 (n = 13)[a]	22.5	2.9	0.8	19	27	8
Personality		2 (n = 15)[b]	22.8	2.0	0.5	20	26	6
Changes		3 (n = 11)[b]	22.5	3.0	0.9	15	26	11
		Total	22.6	2.6	0.4	15	27	12
		(n = 39)						
Lima	36	1 (n = 13)[a]	21.7	3.4	0.9	14	26	12
Vomiting		2 (n = 15)[b]	19.2	4.0	1.0	13	26	13
		3 (n = 11)[b]	24.7	4.4	1.3	17	30	13
		Total	21.6	4.4	0.7	13	30	17
		(n = 39)						

Case	Items	Track	Mean	SD	SEM	Min	Max	Range
Mike	28	1 (n = 13)[a]	17.6	2.5	0.7	13	21	8
Urinary Tract		2 (n = 15)[b]	17.3	2.7	0.7	11	21	10
		3 (n = 11)[b]	19.4	3.4	1.0	14	24	10
		Total (n = 39)	18.0	2.9	0.5	11	24	13
November	23	1 (n = 13)[a]	14.8	2.4	0.7	11	18	7
Cardiac		2 (n = 15)[b]	13.4	2.1	0.6	10	17	7
Counseling		3 (n = 11)[b]	11.6	2.5	0.8	7	16	9
		Total (n = 39)	13.4	2.6	0.4	7	18	11

[a] One group of 14 Standardized Patients (one trained for each case) were assigned to the Blue Track (Track 1; morning session 1)

[b] The second group of 14 Standardized Patients (one trained for each case) were assigned to the two Red Tracks (Track 2; morning session 2) and Track 3 (the lone afternoon session)

Appendix E

Summary Statistics for the Checklist scores (SPs)

Case	Items	Standardized Patient (SP)	Mean	SD	SEM	Min	Max	Range
Alpha	43	Red (n = 26)	27.0	6.0	1.2	12	37	25
Fever		Blue (n = 13)	24.3	4.2	1.2	18	31	13
		Total (n = 39)	26.1	5.6	0.9	12	37	25
Bravo	41	Red (n = 26)	26.9	5.2	1.0	16	36	20
Headache		Blue (n = 13)	22.8	5.6	1.6	12	31	19
		Total (n = 39)	25.5	5.6	0.9	12	36	24
Charlie	28	Red (n = 26)	14.3	2.9	0.6	9	20	11
Infection		Blue (n = 13)	13.5	2.4	0.7	10	17	7
		Total (n = 39)	14.1	2.7	0.4	9	20	11
Delta	13	Red (n = 26)	7.4	2.0	0.4	4	10	6
Informed Consent		Blue (n = 13)	8.7	1.2	0.3	7	10	3
		Total (n = 39)	7.8	1.9	0.3	4	10	6

Case	Items	Standardized Patient	Mean	SD	SEM	Min	Max	Range
Echo	36	Red (n = 26)	22.1	3.3	0.7	15	28	13
Assess Risk		Blue (n = 13)	21.5	3.7	1.0	16	29	13
		Total (n = 39)	21.9	3.4	0.5	15	29	14
Foxtrot	16	Red (n = 26)	8.1	2.5	0.5	3	13	10
Metastasis		Blue (n = 13)	7.6	2.7	0.8	4	12	8
		Total (n = 39)	7.9	2.5	0.4	3	13	10
Golf	33	Red (n = 26)	21.0	2.9	0.6	17	27	10
Breathing		Blue (n = 13)	19.0	2.6	0.7	15	23	8
		Total (n = 39)	20.3	2.9	0.5	15	27	12
Hotel	40	Red (n = 26)	20.7	3.5	0.7	16	28	12
Flu		Blue (n = 13)	21.9	2.7	0.8	17	26	9
		Total (n = 39)	21.1	3.3	0.5	16	28	12
India	32	Red (n = 26)	16.2	2.8	0.6	10	23	13
Fatigue		Blue (n = 13)	21.7	3.7	1.0	14	25	11
		Total (n = 39)	18.0	4.1	0.7	10	25	15

Case	Items	Standardized Patient (SP)	Mean	SD	SEM	Min	Max	Range
Juliet Hand	35	Red (n = 26)	19.0	3.3	0.6	12	26	14
		Blue (n = 13)	21.3	4.1	1.2	12	27	15
		Total (n = 39)	19.8	3.7	0.6	12	27	15
Kilo Personality	35	Red (n = 26)	22.7	2.4	0.5	15	26	11
		Blue (n = 13)	22.5	2.9	0.8	19	27	8
		Total (n = 39)	22.6	2.6	0.4	15	27	12
Lima Vomiting	36	Red (n = 26)	21.5	5.0	1.0	13	30	17
		Blue (n = 13)	21.7	3.4	0.9	14	26	12
		Total (n = 39)	21.6	4.4	0.7	13	30	17
Mike Urinary	28	Red (n = 26)	18.2	3.2	0.6	11	24	13
		Blue (n = 13)	17.6	2.5	0.7	13	21	8
		Total (n = 39)	18.0	2.9	0.5	11	24	13
November Cardiac	23	Red (n = 26)	12.7	2.4	0.5	7	17	10
		Blue (n = 13)	14.8	2.4	0.8	11	18	7
		Total (n = 39)	13.4	2.6	0.4	7	18	11

Appendix F

Summary Statistics for the Global scores (Tracks 1, 2, and 3)

Case	Track	Mean	SD	SEM	Min	Max	Range
Alpha Fever	1 (n = 13)[a]	2.0	.71	.20	1	3	2
	2 (n = 15)[b]	3.7	.82	.21	2	5	3
	3 (n = 11)[b]	3.6	1.4	.43	1	5	4
	Total (n = 39)	3.1	1.2	.20	1	5	4
Bravo Headache	1 (n = 13)[a]	2.5	1.1	0.3	1	4	3
	2 (n = 15)[b]	3.1	1.00	0.3	1	4	3
	3 (n = 11)[b]	3.2	1.3	0.4	1	4	3
	Total (n = 39)	2.9	1.1	0.2	1	4	3
Charlie Infection	1 (n = 13)[a]	2.1	1.0	0.3	1	4	3
	2 (n = 15)[b]	2.9	0.7	0.2	2	4	2
	3 (n = 11)[b]	2.3	1.0	0.3	1	4	3
	Total (n = 39)	2.4	1.0	0.2	1	4	3
Delta Informed Consent	1 (n = 13)[a]	4.2	0.4	0.1	4	5	1
	2 (n = 15)[b]	2.3	1.2	0.3	1	4	3
	3 (n = 11)[b]	4.2	0.6	0.2	3	5	2
	Total (n = 39)	3.5	1.3	0.20	1	5	4

Case	Track	Mean	SD	SEM	Min	Max	Range
Echo	1 (n = 13)[a]	3.4	0.8	0.2	2	4	2
Risk	2 (n = 15)[b]	3.5	0.5	0.1	3	4	1
Assessment	3 (n = 11)[b]	2.8	1.1	0.3	1	5	4
	Total (n = 39)	3.3	0.8	0.1	1	5	4
Foxtrot	1 (n = 13)[a]	1.9	1.3	0.4	1	4	3
Metastasis	2 (n = 15)[b]	3.1	1.2	0.3	1	5	4
	3 (n = 11)[b]	2.9	1.3	0.4	1	4	3
	Total (n = 39)	2.7	1.3	0.2	1	5	4
Golf	1 (n = 13)[a]	2.9	1.0	0.3	1	4	3
Shortness of	2 (n = 15)[b]	3.5	0.5	0.1	3	4	1
Breath	3 (n = 11)[b]	3.7	1.1	0.3	2	5	3
	Total (n = 39)	3.4	0.9	0.1	1	5	4
Hotel	1 (n = 13)[a]	3.4	0.7	0.2	2	4	2
Flu	2 (n = 15)[b]	3.3	0.8	0.2	2	4	2
Symptoms	3 (n = 11)[b]	2.2	1.0	0.3	1	4	3
	Total (n = 39)	3.00	1.0	0.2	1	4	3

Case	Track	Mean	SD	SEM	Min	Max	Range
India	1 (n = 13)[a]	3.4	0.9	0.2	2	5	3
Fatigue	2 (n = 15)[b]	2.7	1.5	0.4	1	5	4
	3 (n = 11)[b]	2.5	1.0	0.3	1	4	3
	Total (n = 39)	2.9	1.2	0.2	1	5	4
Juliet	1 (n = 13)[a]	3.2	1.0	0.3	1	4	3
Hand	2 (n = 15)[b]	3.1	0.7	0.2	2	4	2
Problems	3 (n = 11)[b]	3.00	1.0	0.3	1	5	4
	Total (n = 39)	3.1	0.9	0.1	1	5	4
Kilo	1 (n = 13)[a]	3.2	0.6	0.2	2	4	2
Personality	2 (n = 15)[b]	3.9	0.3	0.1	3	4	1
Changes	3 (n = 11)[b]	3.0	0.6	0.2	2	4	2
	Total (n = 39)	3.4	0.6	0.1	2	4	2
Lima	1 (n = 13)[a]	2.5	1.1	0.3	1	4	3
Vomiting	2 (n = 15)[b]	3.0	1.2	0.3	1	5	4
	3 (n = 11)[b]	3.5	0.7	0.2	2	4	2
	Total (n = 39)	3.0	1.1	0.2	1	5	4

Case	Track	Mean	SD	SEM	Min	Max	Range
Mike	1 (n = 13)[a]	3.8	0.7	0.2	2	5	3
Urinary Tract	2 (n = 15)[b]	3.7	0.6	0.2	2	4	2
	3 (n = 11)[b]	3.4	0.8	0.2	2	4	2
	Total (n = 39)	3.6	0.7	0.1	2	5	3
November	1 (n = 13)[a]	3.6	0.8	0.2	2	5	3
Cardiac	2 (n = 15)[b]	3.9	0.8	0.2	2	5	3
Counseling	3 (n = 11)[b]	3.4	1.1	0.3	1	5	4
	Total (n = 39)	3.6	0.9	0.1	1	5	4

[a] One group of 14 Standardized Patients (one trained for each case) were assigned to the Blue Track (Track 1; morning session 1)

[b] The second group of 14 Standardized Patients (one trained for each case) were assigned to the two Red Tracks (Track 2; morning session 2) and Track 3 (the lone afternoon session)

Appendix G

Summary Statistics for the Global scores (Two SPs)

Case	Standardized Patient (SP)	Mean	SD	SEM	Min	Max	Range
Alpha	Red (n = 26)[a]	3.6	1.1	0.2	1	5	4
Fever	Blue (n = 13)[b]	2.0	0.7	0.2	1	3	2
	Total (n = 39)	3.1	1.2	0.2	1	5	4
Bravo	Red (n = 26)[a]	3.2	1.1	0.2	1	4	3
Headache	Blue (n = 13)[b]	2.5	1.1	0.3	1	4	3
	Total (n = 39)	2.9	1.1	0.2	1	4	3
Charlie	Red (n = 26)[a]	2.6	0.9	0.2	1	4	3
Infection	Blue (n = 13)[b]	2.1	1.0	0.3	1	4	3
	Total (n = 39)	2.4	1.0	0.2	1	4	3
Delta	Red (n = 26)[a]	3.1	1.4	0.3	1	5	4
Informed Consent	Blue (n = 13)[b]	4.2	0.4	0.1	4	5	1
	Total (n = 39)	3.5	1.3	0.2	1	5	4

Case	Standardized Patient	Mean	SD.	SEM	Min	Max	Range
Echo	Red (n = 26)[a]	3.2	0.9	0.2	1	5	4
Assess Risk	Blue (n = 13)[b]	3.4	0.8	0.2	2	4	2
	Total (n = 39)	3.3	0.8	0.1	1	5	4
Foxtrot	Red (n = 26)[a]	3.0	1.2	0.2	1	5	5
Metastasis	Blue (n = 13)[b]	1.9	1.3	0.4	1	4	3
	Total (n = 39)	2.7	1.3	0.2	1	5	4
Golf	Red (n = 26)[a]	3.6	0.8	0.2	2	5	3
Breathing	Blue (n = 13)[b]	2.9	1.0	0.3	1	4	3
	Total (n = 39)	3.4	0.9	0.1	1	5	4
Hotel	Red (n = 26)[a]	2.8	1.0	0.2	1	4	3
Flu	Blue (n = 13)[b]	3.4	0.7	0.2	2	4	2
	Total (n = 39)	3.0	1.0	0.2	1	4	3
India	Red (n = 26)[a]	2.6	1.3	0.3	1	5	4
Fatigue	Blue (n = 13)[b]	3.4	0.9	0.2	2	5	3
	Total (n = 39)	2.9	1.2	0.2	1	5	4

Case	Standardized Patient	Mean	SD	SEM	Min	Max	Range
Juliet	Red (n = 26)[a]	3.0	0.8	0.2	1	5	4
Hand	Blue (n = 13)[b]	3.2	1.0	0.3	1	4	3
	Total (n = 39)	3.1	0.9	0.1	1	5	4
Kilo	Red (n = 26)[a]	3.5	0.7	0.1	2	4	2
Personality	Blue (n = 13)[b]	3.2	0.6	0.2	2	4	2
	Total (n = 39)	3.4	0.6	0.1	2	4	2
Lima	Red (n = 26)[a]	3.2	1.0	0.2	1	5	4
Vomiting	Blue (n = 13)[b]	2.5	1.1	0.3	1	4	3
	Total (n = 39)	3.0	1.1	0.2	1	5	5
Mike	Red (n = 26)[a]	3.5	0.7	0.1	2	4	2
Urinary	Blue (n = 13)[b]	3.8	0.7	0.2	2	5	3
	Total (n = 39)	3.6	0.7	0.1	2	5	3
November	Red (n = 26)[a]	3.7	1.0	0.2	1	5	4
Cardiac	Blue (n = 13)[b]	3.6	0.8	0.2	2	5	3
	Total (n = 39)	3.6	0.9	0.1	1	5	4

Appendix H

Summary Statistics for the Total scores (Tracks 1, 2, and 3)

Case	Total Score	Track	Mean	SD	SEM	Min	Max	Range
Alpha	48	1 (n = 13)[a]	26.3	4.7	1.3	19	34	15
Fever	(43+5)	2 (n = 15)[b]	29.5	6.1	1.6	15	41	26
		3 (n = 11)[b]	32.2	7.6	2.3	14	42	28
		Total (n = 39)	29.2	6.4	1.0	14	42	28
Bravo	46	1 (n = 13)[a]	25.2	6.5	1.8	13	34	21
Headache	(41+5)	2 (n = 15)[b]	30.9	5.2	1.3	22	38	16
		3 (n = 11)[b]	28.8	7.2	2.2	17	40	23
		Total (n = 39)	28.4	6.5	1.1	13	40	27
Charlie	33	1 (n = 13)[a]	15.6	3.2	0.9	11	21	10
Infection	(28+5)	2 (n = 15)[b]	18.1	3.3	0.8	13	24	11
		3 (n = 11)[b]	15.3	3.5	1.1	10	21	11
		Total (n = 39)	16.5	3.5	0.6	10	24	14
Delta	18	1 (n = 13)[a]	12.9	1.5	0.4	11	15	4
Informed	(13+5)	2 (n = 15)[b]	8.7	3.0	0.8	5	14	9
Consent		3 (n = 11)[b]	12.9	1.7	0.5	9	15	6
		Total (n = 39)	11.3	3.0	0.5	5	15	10

Case	Total Score	Track	Mean	SD	SEM	Min	Max	Range
Echo	41	1 (n = 13)[a]	24.9	4.3	1.2	19	33	14
Risk	(36+5)	2 (n = 15)[b]	25.3	4.2	1.1	18	32	14
Assessment		3 (n = 11)[b]	25.3	3.6	1.1	20	33	13
		Total (n = 39)	25.2	3.9	0.6	18	33	15
Foxtrot	21	1 (n = 13)[a]	9.5	3.7	1.0	5	16	11
Metastasis	(16+5)	2 (n = 15)[b]	11.5	2.5	0.6	7	15	8
		3 (n = 11)[b]	10.6	4.2	1.3	4	17	13
		Total (n = 39)	10.6	3.4	0.6	4	17	13
Golf	38	1 (n = 13)[a]	21.9	3.2	0.9	17	27	10
Shortness of	(33+5)	2 (n = 15)[b]	24.3	2.8	0.7	21	29	8
Breath		3 (n = 11)[b]	24.9	4.8	1.4	20	32	12
		Total (n = 39)	23.7	3.6	0.6	17	32	15
Hotel	45	1 (n = 13)[a]	25.3	3.2	0.9	19	29	10
Flu	(40+5)	2 (n = 15)[b]	22.9	3.8	1.0	18	32	14
Symptoms		3 (n = 11)[b]	24.3	4.3	1.3	17	30	13
		Total (n = 39)	24.1	3.8	0.6	17	32	15

Case	Total Score	Track	Mean	SD	SEM	Min	Max	Range
India	37	1 (n = 13)[a]	25.1	4.3	1.2	16	29	13
Fatigue	(32+5)	2 (n = 15)[b]	18.9	4.6	1.2	11	28	17
		3 (n = 11)[b]	18.6	2.6	0.8	14	22	8
		Total	20.9	4.9	0.8	11	29	18
		(n = 39)						
Juliet	40	1 (n = 13)[a]	24.5	4.9	1.4	13	31	18
Hand	(35+5)	2 (n = 15)[b]	22.3	3.5	0.9	15	27	12
Problems		3 (n = 11)[b]	21.7	4.7	1.4	13	31	18
		Total	22.9	4.4	0.7	13	31	18
		(n = 39)						
Kilo	40	1 (n = 13)[a]	25.7	3.4	0.9	21	31	10
Personality	(35+5)	2 (n = 15)[b]	26.7	2.1	0.5	24	30	6
Changes		3 (n = 11)[b]	25.5	3.5	1.0	17	31	14
		Total	26.0	2.9	0.5	17	31	14
		(n = 39)						
Lima	41	1 (n = 13)[a]	24.2	3.8	1.1	16	29	13
Vomiting	(36+5)	2 (n = 15)[b]	22.2	5.0	1.3	14	31	17
		3 (n = 11)[b]	28.2	4.9	1.5	19	34	15
		Total	24.6	5.1	0.8	14	34	20
		(n = 39)						

Case	Total Score	Track	Mean	SD	SEM	Min	Max	Range
Mike	33	1 (n = 13)[a]	21.4	3.1	0.9	15	26	11
Urinary Tract	(28+5)	2 (n = 15)[b]	20.9	3.2	0.8	14	25	11
		3 (n = 11)[b]	22.7	4.1	1.2	17	28	11
		Total	21.6	3.5	0.6	14	28	14
		(n = 39)						
November	28	1 (n = 13)[a]	18.4	3.1	0.9	13	22	9
Cardiac	(23+5)	2 (n = 15)[b]	17.3	2.7	0.7	14	22	8
Counseling		3 (n = 11)[b]	15.0	3.5	1.1	8	21	13
		Total	17.0	3.3	0.5	8	22	14
		(n = 39)						

[a] One group of 14 Standardized Patients (one trained for each case) were assigned to the Blue Track (Track 1; morning session 1)

[b] The second group of 14 Standardized Patients (one trained for each case) were assigned to the two Red Tracks (Track 2; morning session 2) and Track 3 (the lone afternoon session)

Appendix I

Summary Statistics for the Total scores (SPs)

Case	Items	Standardized Patient (SP)	Mean	SD	SEM	Min	Max	Range
Alpha Fever	48	Red (n = 26)[a]	30.7	6.8	1.3	14	42	28
		Blue (n = 13)[b]	26.3	4.7	1.3	19	34	15
		Total (n = 39)	29.2	6.4	1.0	14	42	28
Bravo Headache	46	Red (n = 26)[a]	30.0	6.1	1.2	17	40	23
		Blue (n = 13)[b]	25.2	6.5	1.8	13	34	21
		Total (n = 39)	28.4	6.5	1.1	13	40	27
Charlie Infection	33	Red (n = 26)[a]	16.9	3.6	0.7	10	24	14
		Blue (n = 13)[b]	15.6	3.2	0.9	11	21	10
		Total (n = 39)	16.5	3.5	0.6	10	24	14
Delta Informed Consent	18	Red (n = 26)[a]	10.5	3.3	0.5	5	15	10
		Blue (n = 13)[b]	12.9	1.5	0.4	11	15	4
		Total (n = 39)	11.3	3.0	0.5	5	15	10

Case	Items	Standardized Patient	Mean	SD	SEM	Min	Max	Range
Echo	41	Red (n = 26)[a]	25.3	3.9	0.8	18	33	15
Assess Risk		Blue (n = 13)[b]	24.9	4.3	1.2	19	33	14
		Total (n = 39)	25.2	3.9	0.6	18	33	15
Foxtrot	21	Red (n = 26)[a]	11.1	3.2	0.6	4	17	13
Metastasis		Blue (n = 13)[b]	9.5	3.7	1.0	5	16	11
		Total (n = 39)	10.6	3.4	0.6	4	17	13
Golf	38	Red (n = 26)[a]	24.5	3.6	0.7	20	32	12
Breathing		Blue (n = 13)[b]	21.9	3.2	0.9	17	27	10
		Total (n = 39)	23.7	3.6	0.6	17	32	15
Hotel	45	Red (n = 26)[a]	23.5	3.9	0. 8	17	32	15
Flu		Blue (n = 13)[b]	25.3	3.2	0.9	19	29	10
		Total (n = 39)	24.1	3. 8	0.6	17	32	15
India	37	Red (n = 26)[a]	18.8	3.8	0.8	11	28	17
Fatigue		Blue (n = 13)[b]	25.1	4.3	1.2	16	29	13
		Total (n = 39)	20.9	4.9	0.8	11	29	18

Case	Items	Standardized Patient	Mean	SD	SEM	Min	Max	Range
Juliet	40	Red (n = 26)[a]	22.1	3.9	0.8	13	31	18
Hand		Blue (n = 13)[b]	24.5	4.9	1.4	13	31	18
		Total (n = 39)	22.9	4.4	0.7	13	31	18
Kilo	40	Red (n = 26)[a]	26.2	2.7	0.5	17	30	13
Personality		Blue (n = 13)[b]	25.7	3.4	0.9	21	31	10
		Total (n = 39)	26.0	2.9	0.5	17	31	14
Lima	41	Red (n = 26)[a]	24.7	5.7	1.1	14	34	20
Vomiting		Blue (n = 13)[b]	24.2	3.8	1.1	16	29	13
		Total (n = 39)	24.6	5.1	0.8	14	34	20
Mike	33	Red (n = 26)[a]	21.7	3.7	0.7	14	28	14
Urinary		Blue (n = 13)[b]	21.4	3.1	0.9	15	26	11
		Total (n = 39)	21.6	3.4	0.6	14	28	14
November	28	Red (n = 26)[a]	16.3	3.2	0.6	8	22	14
Cardiac		Blue (n = 13)[b]	18.4	3.1	0.9	13	22	9
		Total (n = 39)	17.0	3.3	0.5	8	22	14

Appendix J

ANOVA results for the Checklist Scores (Tracks 1, 2, and 3)

Case	Case Focus	df	F	Sig.
Alpha	HT and PE* (Fever)	2,36	1.930	0.160
Bravo	HT and PE* (Headache)	2,36	3.00	0.062
Charlie	HT and PE* (Infection)	2,36	2.737	0.078
Delta	Informed Consent (Risk Assessment)	2,36	10.993	0.001*
Echo	Counseling (Breast cancer)	2,36	0.217	0.806
Foxtrot	Information Sharing (Cancer metastasis)	2,36	0.315	0.732
Golf	HT and PE* (Shortness of Breath)	2,36	2.142	0.132
Hotel	HT and PE* (Flu Symptoms)	2,36	2.585	0.089
India	HT and PE* (Fatigue)	2,36	12.795	0.001*
Juliet	HT and PE* (Problems with hand)	2,36	1.788	0.182
Kilo	HT and PE* (Personality Changes)	2,36	0.079	0.924
Lima	HT and PE* (Vomiting)	2,36	6.313	0.004*
Mike	HT and PE* (Urinary Tract)	2,36	1.859	0.171
November	Information Sharing (Cardiac counseling)	2,36	5.362	0.009*

Note: HT and PE: History Taking and Physical Examination

* statistical significance at $p < 0.05$

Appendix K

ANOVA results for the Checklist Scores (SPs)

Case	Case Focus	df	F	Sig.
Alpha	HT and PE* (Fever)	1,37	2.163	0.150
Bravo	HT and PE* (Headache)	1,37	5.030	0.031*
Charlie	HT and PE* (Infection)	1,37	0.676	0.416
Delta	Informed Consent (Risk Assessment)	1,37	4.772	0.035*
Echo	Counseling (Breast cancer)	1,37	0.213	0.647
Foxtrot	Information Sharing (Cancer metastasis)	1,37	0.284	0.598
Golf	HT and PE* (Shortness of Breath)	1,37	4.270	0.046*
Hotel	HT and PE* (Flu Symptoms)	1,37	1.250	0.271
India	HT and PE* (Fatigue)	1,37	26.301	0.001*
Juliet	HT and PE* (Problems with hand)	1,37	3.515	0.069
Kilo	HT and PE* (Personality Changes)	1,37	0.048	0.828
Lima	HT and PE* (Vomiting)	1,37	0.010	0.920
Mike	HT and PE* (Urinary Tract)	1,37	0.288	0.595
November	Information Sharing (Cardiac counseling)	1,37	6.637	0.014*

Note: HT and PE: History Taking and Physical Examination

* statistical significance at $p < 0.05$

Appendix L

ANOVA results for the Global Scores Based (Tracks 1, 2, and 3)

Case	Case Focus	df	F	Sig.
Alpha	HT and PE* (Fever)	2,36	11.335	.0001*
Bravo	HT and PE* (Headache)	2,36	1.682	.200
Charlie	HT and PE* (Infection)	2,36	2.773	.076
Delta	Informed Consent (Risk Assessment)	2,36	23.316	.0001*
Echo	Counseling (Breast cancer)	2,36	2.767	.076
Foxtrot	Information Sharing (Cancer metastasis)	2,36	3.596	.038*
Golf	HT and PE* (Shortness of Breath)	2,36	2.774	.076
Hotel	HT and PE* (Flu Symptoms)	2,36	7.881	.001*
India	HT and PE* (Fatigue)	2,36	2.091	.138
Juliet	HT and PE* (Problems with hand)	2,36	.090	.914
Kilo	HT and PE* (Personality Changes)	2,36	12.366	.0001*
Lima	HT and PE* (Vomiting)	2,36	2.26	.119
Mike	HT and PE* (Urinary Tract)	2,36	1.034	.366
November	Information Sharing (Cardiac counseling)	2,36	.993	.381

Note: HT and PE: History Taking and Physical Examination

* statistical significance at $p < 0.05$

Appendix M

ANOVA results for the Global Scores Based (SPs)

Case	Case Focus	df	F	Sig.
Alpha	HT and PE* (Fever)	1,37	23.145	0.001*
Bravo	HT and PE* (Headache)	1,37	3.445	0.071
Charlie	HT and PE* (Infection)	1,37	2.811	0.102
Delta	Informed Consent (Risk Assessment)	1,37	8.512	0.006*
Echo	Counseling (Breast cancer)	1,37	0.295	0.590
Foxtrot	Information Sharing (Cancer metastasis)	1,37	7.139	0.011*
Golf	HT and PE* (Shortness of Breath)	1,37	5.027	0.031*
Hotel	HT and PE* (Flu Symptoms)	1,37	3.430	0.072
India	HT and PE* (Fatigue)	1,37	4.068	0.051
Juliet	HT and PE* (Problems with hand)	1,37	0.149	0.702
Kilo	HT and PE* (Personality Changes)	1,37	2.056	0.160
Lima	HT and PE* (Vomiting)	1,37	3.322	0.076
Mike	HT and PE* (Urinary Tract)	1,37	0.910	0.346
November	Information Sharing (Cardiac counseling)	1,37	0.015	0.902

Note: HT and PE: History Taking and Physical Examination

* statistical significance at $p < 0.05$

Appendix N

ANOVA results for the Total Scores (Tracks 1, 2, and 3)

Case	Case Focus	df	F	Sig.
Alpha	HT and PE* (Fever)	2,36	2.752	.077
Bravo	HT and PE* (Headache)	2,36	2.869	.070
Charlie	HT and PE* (Infection)	2,36	3.042	.060
Delta	Informed Consent (Risk Assessment)	2,36	16.385	.0001*
Echo	Counseling (Breast cancer)	2,36	0.040	.961
Foxtrot	Information Sharing (Cancer metastasis)	2,36	1.107	.341
Golf	HT and PE* (Shortness of Breath)	2,36	2.551	.092
Hotel	HT and PE* (Flu Symptoms)	2,36	1.418	.255
India	HT and PE* (Fatigue)	2,36	10.544	.0001*
Juliet	HT and PE* (Problems with hand)	2,36	1.349	.272
Kilo	HT and PE* (Personality Changes)	2,36	.714	.497
Lima	HT and PE* (Vomiting)	2,36	5.413	.009*
Mike	HT and PE* (Urinary Tract)	2,36	.889	.420
November	Information Sharing (Cardiac counseling)	2,36	3.684	.035*

Note: HT and PE: History Taking and Physical Examination

* statistical significance at $p < 0.05$

Appendix O

ANOVA results for the Total Scores (SPs)

Case	Case Focus	df	F	Sig.
Alpha	HT and PE* (Fever)	1,37	4.306	0.045*
Bravo	HT and PE* (Headache)	1,37	5.099	0.030*
Charlie	HT and PE* (Infection)	1,37	1.227	0.275
Delta	Informed Consent (Risk Assessment)	1,37	6.534	0.015*
Echo	Counseling (Breast cancer)	1,37	0.081	0.778
Foxtrot	Information Sharing (Cancer metastasis)	1,37	1.872	0.179
Golf	HT and PE* (Shortness of Breath)	1,37	4.992	0.032*
Hotel	HT and PE* (Flu Symptoms)	1,37	2.034	0.162
India	HT and PE* (Fatigue)	1,37	21.641	0.001*
Juliet	HT and PE* (Problems with hand)	1,37	2.639	0.113
Kilo	HT and PE* (Personality Changes)	1,37	0.245	0.623
Lima	HT and PE* (Vomiting)	1,37	0.081	0.778
Mike	HT and PE* (Urinary Tract)	1,37	0.067	0.797
November	Information Sharing (Cardiac counseling)	1,37	3.692	0.062

Note: HT and PE: History Taking and Physical Examination

* statistical significance at $p < 0.05$

Appendix P

Physician rater global score and pass/fail status for each case.

			Global Rating		
Case	Poor	Borderline Fail	Borderline Pass	Good	Excellent
Alpha	5	8 (6)[b]	10 (2)[c]	11	5
Bravo	7 (3)[a]	5 (3)[b]	11	16	0
Charlie	8	11 (3)[b]	15 (3)[c]	5	0
Delta	6	2 (1)[b]	5	20	6
Echo	1	5 (2)[b]	16(6)[c]	16	1
Foxtrot	10 (3)[a]	9 (6)[b]	7	10	3
Golf	1 (1)[a]	4	18(1)[c]	12	4
Hotel	3	8 (3)[b]	14(5)[c]	14	0
India	8 (2)[a]	5 (3)[b]	14	9	3
Juliet	2	6 (2)[b]	19(8)[c]	11	1
Kilo	0	3 (1)[b]	16	20	0
Lima	5 (1)[a]	7 (2)[b]	12(2)[c]	14	1
Mike	0	4	8(3)[c]	26	1
November	1	3 (2)[b]	10(1)[c]	20	5
Total Interactions (Total [a, b, c])	57 (10[a])	80 (34[b])	175 (31[c])	204	30
Percent [a, b, c]	17.5%	42.5%	17.7%	0%	0%

[a] Global score = 1 (Poor); Candidate still passed the case

[b] Global score = 2 (Borderline Fail); Candidate still passed the case

[c] Global score = 3 (Borderline Pass); Candidate still failed the case

Appendix Q

Communication Checklist (Physician raters' version)

EXAMINER'S COMMUNICATION CHECKLIST

Simulated Patient ID: ______________________

Registrant ID: ______________________

INSTRUCTIONS: Please place an X in the bubble which conveys your feelings about this doctor. Add up all subtotals and write the totals in the appropriate boxes below.

	Strongly Disagree 1	Disagree 2	Not Sure 3	Agree 4	Strongly Agree 5
1. The doctor wanted to understand how the patient saw things.	○	○	○	○	○
2. The doctor usually sensed or realized what the patient was feeling.	○	○	○	○	○
3. The doctor just took no notice of some things that the patient thought or felt.	○	○	○	○	○
4. The doctor's response to the patient was usually so fixed & automatic that the patient didn't really get through to him/her.	○	○	○	○	○
5. The doctor treated the patient with respect & courtesy.	○	○	○	○	○
6. The patient was able to explain his/her problem to the doctor as fully as needed.	○	○	○	○	○
7. The doctor explained things to the patient so that they know what may be the matter with them.	○	○	○	○	○
8. The doctor explained what treatment, tests or other follow-up is going to happen.	○	○	○	○	○
9. The doctor gave the patient the opportunity to express his/her feelings or ideas in planning treatment, tests or follow-up.	○	○	○	○	○
10. The doctor gave the patient the opportunity to ask questions.	○	○	○	○	○
11. The doctor used understandable and non-technical language.	○	○	○	○	○
12. The doctor was careful and thorough.	○	○	○	○	○
13. The patient feels satisfied with the medical care that he/she received.	○	○	○	○	○
Subtotals					

TOTALS IPS

Appendix R

Generalizability analyses of the Communication Checklist Items

Communication Item	SP rater		Physician rater	
	Facet	% Variance	Facet	% Variance
Q1. The doctor wanted to understand how the patient saw things.	p	3.1	p	3.8
	c	23.2	c	0
	p x c	52.5	p x c	51.2
	sp:c	21.2	r in c	30.1
			c x r in c	14.9
Q2. The doctor usually sensed or realized what the patient was feeling	p	6.3	p	6.5
	c	11.3	c	0
	p x c	55.4	p x c	55.2
	sp:c	27.0	r in c	28.0
			c x r in c	10.3
Q3. The doctor just took no notice of some things that the patient thought or felt.	p	3.1	p	1.2
	c	7.5	c	0.0
	p x c	71.1	p x c	59.0
	sp:c	18.2	r in c	35.5
			c x r in c	4.2

Communication Item	SP rater		Physician rater	
Q4. The doctor's response to the patient was usually so fixed and automatic that the patient didn't really get through to him/her.	p	7.0	p	6.6
	c	5.3	c	0.0
	p x c	70.0	p x c	56.4
	sp:c	17.7	r in c	29.2
			c x r in c	7.7
Q5. The doctor treated the patient with respect and courtesy.	p	3.4	p	4.2
	c	0.0	c	0.0
	p x c	62.8	p x c	44.4
	sp:c	33.8	r in c	51.4
			c x r in c	0.0
Q6. The patient was able to explain his/her problem to the doctor as fully as needed.	p	1.9	p	2.4
	c	14.9	c	0.0
	p x c	54.3	p x c	70.6
	sp:c	29.2	r:c	24.7
			c x r:c	2.3

Communication Item	SP Rater		Physician Rater	
Q7. The doctor explained things to the patient so that they know what may be the matter with them.	p	2.4	p	2.4
	c	2.9	c	2.0
	p x c	68.6	p x c	86.1
	sp:c	26.1	r:c	3.0
			c x r:c	6.5
Q8. The doctor explained what treatment, tests, or other follow-up is going to happen.	p	2.4	p	3.6
	c	12.7	c	2.7
	p x c	67.1	p x c	83.8
	sp:c	17.8	r:c	3.4
			c x r:c	6.6
Q9. The doctor gave the patient the opportunity to express his/her feelings or ideas in planning treatment, tests, or follow-up.	p	4.0	p	5.0
	c	17.3	c	0.3
	p x c	51.3	p x c	57.2
	sp:c	27.3	r:c	32.7
			c x r:c	4.9

Communication Item	SP Rater		Physician Rater	
Q10. The doctor gave the patient the opportunity to ask questions.	p	5.7	p	5.7
	c	11.7	c	9.2
	p x c	55.5	p x c	60.3
	sp:c	27.2	r:c	24.7
			c x r:c	0.0
Q11. The doctor used understandable and non-technical language.	p	5.4	p	7.9
	c	0.0	c	0.0
	p x c	58.1	p x c	62.6
	sp:c	36.5	r:c	29.5
			c x r:c	0.0
Q12. The doctor was careful and thorough.	p	5.8	p	8.8
	c	11.4	c	0.0
	p x c	60.3	p x c	59.6
	sp:c	22.6	r:c	18.3
			c x r:c	13.2

Communication Item	SP Rater		Physician Rater	
Q13. The patient feels satisfied with the medical care that he/she received.	p	6.9	p	9.1
	c	1.1	c	0.0
	p x c	54.5	p x c	57.5
	sp:c	37.5	r:c	25.8
			c x r:c	8.3

Appendix S

A comparison of communication item variance between SP and physician raters

Facet	SPs		Physician Raters	
p*	Item 2	Sensed or realized feelings	Item 2	Sensed or realized feelings
	Item 4	Fixed and automatic response	Item 4	Fixed and automatic response
			Item 11	Used understandable/non-technical language
			Item 12	Physician was careful and thorough
	Item 13	Patient satisfied with care	Item 13	Patient satisfied with care
c**	Item 1	Understand how patient understood things		
	Item 2	Sensed or realized feelings		
	Item 6	Able to explain problem		
	Item 8	Doctor explained tests, etc.		
	Item 9	Doctor allowed patient opinion on tests, etc.		
	Item 10	Allowed patient to ask questions		
	Item 11	Used understandable/non-technical language		

Facet	SPs		Physician Raters	
p x c ***	Item 3	Doctor took no notice of feelings		
	Item 4	Fixed and automatic response		
	Item5	Showed respect and courtesy		
			Item 6	Able to explain problem
	Item 7	What is the matter with patient	Item 7	What is the matter with patient
	Item 8	Doctor explained tests, etc.	Item 8	Doctor explained tests, etc.
			Item 10	Allowed patient to ask questions
			Item 11	Used understandable/non-technical language
	Item 12	Doctor was careful and thorough		

Note: p = participants, c = cases, p x c = participants crossed with cases, n = nested into case (for SPs) and assigned to case for physician raters

p* - items with a percent variance above 6.0%

c** - items with a percent variance above 10.0%

p x c*** - items with a percent variance above 60%

Appendix T

A comparison between nested SPs and assigned physician raters on communication items with a variance above 20%

SPs nested		Physician Raters Assigned	
Item 1	Understood how patient (pt.) saw things	Item 1	Understood how patient (pt.) saw things
Item 2	Sensed what pt. was feeling	Item 2	Sensed what pt. was feeling
		Item 3	Doctor took no notice of what pt. said
		Item 4	Fixed and automatic response
Item 5	Respect and courtesy to pt.	Item 5	Respect and courtesy to pt.
Item 6	Pt. able to explain problem	Item 6	Pt. able to explain problem
Item 7	Explained what might be wrong		
Item 9	Pt. allowed input on treatment, tests, etc.	Item 9	Pt. allowed input on treatment, tests, etc.
Item 10	Patient could ask questions	Item 10	Pt. could ask questions
Item 11	Doctor used non-tech language	Item 11	Doctor used non-tech language
Item 12	Doctor careful and thorough		
Item 13	Pt. was satisfied with care	Item 13	Pt. was satisfied with care

Note: Item 8 (The doctor explain treatment and tests to patient) below 20% variance

Appendix U

ANOVA results comparing SP communication scores by case

Station	Comparison by Track or SP pairing	df	F	Sig.
Alpha	Track 1 (Blue), 2 (Red), and 3 (Red)	2,36	1.228	.305
(Fever)	SP pairing (Blue and Red)	1,37	.696	.410
Bravo	Track 1 (Blue), 2 (Red), and 3 (Red)	2,29	0.104	.901
(Headache)	SP pairing (Blue and Red)	1,30	.009	.926
Charlie	Track 1 (Blue), 2 (Red), and 3 (Red)	2,31	5.966	.006*
(Infection)	SP pairing (Blue and Red)	1.32	11.745	.002*
Delta	Track 1 (Blue), 2 (Red), and 3 (Red)	2,36	7.610	.002*
(Informed Consent)	SP pairing (Blue and Red)	1,37	13.634	.001*
Echo	Track 1 (Blue), 2 (Red), and 3 (Red)	2,34	0.471	.628
(Risk Assessment)	SP pairing (Blue and Red)	1.35	0.842	.365
Foxtrot	Track 1 (Blue), 2 (Red), and 3 (Red)	2,35	1.541	.228
(Metastasis Cancer)	SP pairing (Blue and Red)	1,36	3.154	.084
Golf	Track 1 (Blue), 2 (Red), and 3 (Red)	2,36	6.838	.003*
(Shortness of Breath)	SP pairing (Blue and Red)	1,37	10.198	.003*
Hotel	Track 1 (Blue), 2 (Red), and 3 (Red)	2,35	8.403	.001*
(Flu Symptoms)	SP pairing (Blue and Red)	1,36	12.790	.001*
India	Track 1 (Blue), 2 (Red), and 3 (Red)	2,32	69.551	.0001*
(Fatigue)	SP pairing (Blue and Red)	1,33	135.05	.0001*
Juliet	Track 1 (Blue), 2 (Red), and 3 (Red)	2,34	10.671	.0001*
(Hand Problem)	SP pairing (Blue and Red)	1,35	21.738	.0001*

Station	Comparison by Track or SP pairing	df	F	Sig.
Kilo	Track 1 (Blue), 2 (Red), and 3 (Red)	2,35	6.982	.003*
(Personality Change)	SP pairing (Blue and Red)	1,36	1.392	.246
Lima	Track 1 (Blue), 2 (Red), and 3 (Red)	2,36	15.287	.0001*
(Vomiting)	SP pairing (Blue and Red)	1,37	31.319	.0001*
Mike	Track 1 (Blue), 2 (Red), and 3 (Red)	2,36	27.421	.0001*
(Urinary Tract)	SP pairing (Blue and Red)	1,37	55.701	.0001*
November	Track 1 (Blue), 2 (Red), and 3 (Red)	2,36	2.301	.115
(Cardiac Counselling)	SP pairing (Blue and Red)	1,37	4.693	.037*

Appendix V

ANOVA results comparing SP communication scores by checklist item

Station	Comparison by Track or SP pairing	df	F	Sig.
Item 1	Track 1 (Blue), 2 (Red), and 3 (Red)	2,34	1.325	.279
(Understand)	SP pairing (Blue and Red)	1,35	1.852	.182
Item 2	Track 1 (Blue), 2 (Red), and 3 (Red)	2,35	2.151	.131
(Sensed feeling)	SP pairing (Blue and Red)	1,36	3.872	.057
Item 3	Track 1 (Blue), 2 (Red), and 3 (Red)	2,33	2.908	.069
(No notice)	SP pairing (Blue and Red)	1.34	5.307	.027*
Item 4	Track 1 (Blue), 2 (Red), and 3 (Red)	2,31	0.247	.783
(Fixed/Automatic)	SP pairing (Blue and Red)	1,32	0.103	.750
Item 5	Track 1 (Blue), 2 (Red), and 3 (Red)	2,32	0.113	.893
(Respect)	SP pairing (Blue and Red)	1,33	0.186	.669
Item 6	Track 1 (Blue), 2 (Red), and 3 (Red)	2,31	3.384	.047*
(Explain problem)	SP pairing (Blue and Red)	1,32	4.962	.033*
Item 7	Track 1 (Blue), 2 (Red), and 3 (Red)	2,30	4.688	.017*
(What is wrong)	SP pairing (Blue and Red)	1,31	8.303	.007*
Item 8	Track 1 (Blue), 2 (Red), and 3 (Red)	2,32	2.413	.106
(Treatment/Tests)	SP pairing (Blue and Red)	1,33	3.735	.062
Item 9	Track 1 (Blue), 2 (Red), and 3 (Red)	2,31	2.427	.105
(Input)	SP pairing (Blue and Red)	1,32	3.257	.081
Item 10	Track 1 (Blue), 2 (Red), and 3 (Red)	2,28	2.452	.104
(Ask questions)	SP pairing (Blue and Red)	1,29	4.791	.037*

Station	Comparison by Track or SP pairing	df	F	Sig.
Item 11	Track 1 (Blue), 2 (Red), and 3 (Red)	2,32	0.027	.974
(Non-technical)	SP pairing (Blue and Red)	1,33	0.027	.871
Item 12	Track 1 (Blue), 2 (Red), and 3 (Red)	2,33	1.67	.847
(Careful thorough)	SP pairing (Blue and Red)	1,34	0.013	.911
Item 13	Track 1 (Blue), 2 (Red), and 3 (Red)	2,33	0.713	.497
(Satisfied with care)	SP pairing (Blue and Red)	1,34	1.101	.301

Appendix W

ANOVA results comparing Physician rater communication scores by case

Station	Comparison by Track or SP pairing	df	F	Sig.
Alpha (Fever)	Track 1 (am), 2 (am), and 3 (pm)	2,36	12.874	.0001*
Bravo (Headache)	Track 1 (am), 2 (am), and 3 (pm)	2,36	11.525	.0001*
Charlie (Infection)	Track 1 (am), 2 (am), and 3 (pm)	2,36	0.322	.727
Delta (Informed Consent)	Track 1 (am), 2 (am), and 3 (pm)	2,36	65.073	.0001*
Echo (Risk Assessment)	Track 1 (am), 2 (am), and 3 (pm)	2,36	4.081	.025*
Foxtrot (Metastasis Cancer)	Track 1 (am), 2 (am), and 3 (pm)	2,36	2.327	.112
Golf (Shortness of Breath)	Track 1 (am), 2 (am), and 3 (pm)	2,36	18.436	.0001*
Hotel (Flu Symptoms)	Track 1 (am), 2 (am), and 3 (pm)	2,36	14.224	.0001*
India (Fatigue)	Track 1 (am), 2 (am), and 3 (pm)	2,35	8.871	.001*
Juliet (Hand Problem)	Track 1 (am), 2 (am), and 3 (pm)	2,34	1.967	.156

Station	Comparison by Track or SP pairing	df	F	Sig.
Kilo (Personality Change)	Track 1 (am), 2 (am), and 3 (pm)	2,35	3.980	.028*
Lima (Vomiting)	Track 1 (am), 2 (am), and 3 (pm)	2,36	16.467	.0001*
Mike (Urinary Tract)	Track 1 (am), 2 (am), and 3 (pm)	2,36	1.466	.244
November (Cardiac Counselling)	Track 1 (am), 2 (am), and 3 (pm)	2,35	0.994	.380

Appendix X

ANOVA results comparing Physician raters communication scores by checklist item

Station	Comparison by Track or SP pairing	df	F	Sig.
Item 1 (Understand)	Track 1 (am), 2 (am), and 3 (pm)	2,36	7.770	.002*
Item 2 (Sensed feeling)	Track 1 (am), 2 (am), and 3 (pm)	2,36	3.031	.061
Item 3 (No notice)	Track 1 (am), 2 (am), and 3 (pm)	2,36	3.189	.053
Item 4 (Fixed/Automatic)	Track 1 (am), 2 (am), and 3 (pm)	2,36	1.528	.231
Item 5 (Respect)	Track 1 (am), 2 (am), and 3 (pm)	2,35	0.515	.602
Item 6 (Explain problem)	Track 1 (am), 2 (am), and 3 (pm)	2,36	4.694	.015*
Item 7 (What is wrong)	Track 1 (am), 2 (am), and 3 (pm)	2,36	2.919	.067
Item 8 (Treatment/Tests)	Track 1 (am), 2 (am), and 3 (pm)	2,34	1.230	.305
Item 9 (Input)	Track 1 (am), 2 (am), and 3 (pm)	2,35	8.207	.001*
Item 10 (Ask questions)	Track 1 (am), 2 (am), and 3 (pm)	2,36	5.606	.008*

Station	Comparison by Track or SP pairing	df	F	Sig.
Item 11 (Non-technical)	Track 1 (am), 2 (am), and 3 (pm)	2,35	0.113	.893
Item 12 (Careful thorough)	Track 1 (am), 2 (am), and 3 (pm)	2,36	1.088	.348
Item 13 (Satisfied with care)	Track 1 (am), 2 (am), and 3 (pm)	2,34	0.952	.396