2009

Psychometrics of OSCE standardized patient measurements

Frederick R. B Stilson
University of South Florida

Psychometrics of OSCE Standardized Patient Measurements

by

Frederick. R. B. Stilson

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctorate of Philosophy
Department of Psychology
College of Arts and Sciences
University of South Florida

Major Professor: Michael T. Brannick, Ph.D.
Walter C. Borman, Ph.D.
Michael D. Coovert, Ph.D.
Dawn M. Schocken, Ph.D.c.
Joseph A. Vandello, Ph. D.

Date of Approval:
May 9, 2008

Keywords: reliability, validity, assessment centers, interdisciplinary, medicine

Dedication

This dissertation is dedicated to my loving wife Katie, and our dogs Texas and Arrow.

Acknowledgements

I would like to thank my wife Katie for support and patience and also for her assistance with rating videos. Additionally, I would like to thank Magda Francois, Christopher Salas, and Marisa Serafini for their assistance with rating videos. A big thank you goes out to Matt Tuttle for his assistance with software and factor analysis. Dan Ispas assisted me with computerizing the rating process which saved me countless hours. I would like to thank Dawn Schocken, from my committee, for allowing us into the medical department and assisting us all along the way any time questions or problems arose. Laura Fowler Pierce assisted me by taking care of the many forms needed for a successful dissertation defense while I was out of state. Martha Lakis and Nicole Barratt were invaluable with helping me put my training program together. Additionally, gratitude is in order for my committee chair, Robert Dedrick, and my other committee members, Michael Coovert, Walter Borman, and Joseph Vandello. I would like to thank my mom who always reminded me that there was a light at the end of the tunnel. Lastly, I would like to thank Mike Brannick for making this dissertation possible and keeping me motivated along the way and provided very useful feedback in a timely manner and always had time to meet with me no matter how full his schedule was during the several semesters I worked on this dissertation.

Table of Contents

List of Tables

Psychometrics of OSCE Standardized Patient Measurements

Frederick. R. B. Stilson

ABSTRACT

This study examined the reliability and validity of scores taken from a series of four task simulations used to evaluate medical students. The four role-play exercises represented two different cases or scripts, yielding two pairs of exercises that are considered alternate forms. The design allowed examining what is essentially the ceiling for reliability and validity of ratings taken in such role plays. A multitrait-multimethod (MTMM) matrix was computed with exercises as methods and competencies (history taking, clinical skills, and communication) as traits. The results within alternate forms (within cases) were then used as a baseline to evaluate the reliability and validity of scores between the alternate forms (between cases). There was much less of an exercise effect (method variance, monomethod bias) in this study than is typically found in MTMM matrices for performance measurement. However, the convergent validity of the dimensions across exercises was weak both within and between cases. The study also examined the reliability of ratings by training raters to watch video recordings of the same four exercises who then complete the same forms used by the standardized patients. Generalizability analysis was used to compute variance components for case, station, rater, and ratee (medical student), which allowed the computation of reliability estimates for multiple designs. Both the generalizability analysis and the MTMM analysis indicated that rather long examinations (approximately 20 to 40 exercises) would be needed to create reliable examination scores for this population of examinees. Additionally, interjudge agreement was better for more objective dimensions (history taking, physical examination) than for the more subjective dimension (communication).

Chapter One

Introduction

An Introduction to Measurement Properties

Clinical competence in medicine is multidimensional, and many different methods have been devised to assess such competence (Norman, 1985). Methods for measuring clinical competence have included direct observation of actual medical care delivery, oral examinations, written examinations, global rating scales, medical records reviews, patient management problems, computer simulations, and simulated patients (Norman, 1985). A detailed description and evaluation of all such techniques is beyond the scope of this paper, which focuses on task simulations.

The Objective Structured Clinical Examination (OSCE, “AH-skee”) is a collection of task simulations used to evaluate the competence of medical students in the diagnosis and treatment of patients (Newble, 2004). The typical OSCE is a high-stakes examination; failure to perform well can keep a medical student from obtaining a license to practice medicine. In the OSCE role-play, actors (called standardized patients) are used to simulate patients according to a pre-determined script (called a case). The medical student examines the standardized patient during the role-play exercise. Often the role-play exercise is observed by a medical faculty member. At the end of the simulation, the student usually answers a series of questions concerning diagnosis and treatment of the standardized patient, and often the standardized patient and/or faculty member complete an evaluation of the student’s performance during the role-play. The student’s answers to questions subsequent to the role-play may also be scored. Typically, the OSCE will be comprised of several different exercises that are completed at different stations, including encounters with multiple standardized patients.

The point of the OSCE is to demonstrate competence in clinical settings, so that procedural knowledge can be assessed. Although multiple choice tests (paper-and-pencil or computer-based) are well designed for assessing factual (declarative) knowledge, they are widely believed to be deficient at assessing skill in actually carrying out tasks (procedural knowledge). Multiple choice tests also lack in the ability to assess how physicians deal with people as social beings rather than simply as biological mechanisms to be fixed when broken. Standardized, rather than actual, patients are used so that each student faces essentially the same situation (patient) and because using real patients poses several problems, such as patients' potential lack of stamina, acting ability, and desire to participate in such evaluations.

Despite the obvious advantages provided by using the role-play exercise to assess clinical skills, there are also potential problems with the approach. Unlike the multiple choice test, where the stimulus (test item) is identical for all practical purposes across examinees, the interaction between two people (physician and patient) cannot be scripted entirely, and so such interactions would be expected to vary, even on repeated testing with the same physician and standardized patient. Because the physician is not scripted, some improvisation by the standardized patient is always required. Even when two standardized patients are given the same script and told to portray the same case, they will differ in many ways and may not provide an equivalent stimulus (consider age, race and sex effects of the standardized patient, for example). There is often human judgment associated with the evaluation of the performance (i.e., the standardized patient, and/or the faculty member will make a rating of the student's performance). Unlike a paper-and-pencil item, where everyone can agree that the student chose a specific letter in response to the item (e.g., the student chose "a," which is the keyed response, so they get a point credit) there will usually be variance in the performance ratings that is associated with the judge (faculty or standardized patient as a rater), particularly for dimensions such as communication, which require the judge to make an essentially qualitative assessment. Therefore, the role-play exercise provides benefits for assessment, but also appears to be subject to measurement problems that can impact the reliability and validity of the scores from the assessment.

The purpose of this dissertation is to explore some of the sources of variance in performance measures that result from OSCE role-play evaluations. By understanding the nature and magnitude of sources of variance in evaluations of role-plays with standardized patients, we will be able to understand better the reliability and validity of such evaluations. The project has both practical and theoretical aims. Some practical applications of the results of the dissertation concern the design and administration of exercises in the OSCE. That is, the results can be used to spot potential problems in the interpretations of OSCE scores, and thus show where best to aim efforts at improving the evaluation process. Because the OSCE is a high-stakes examination tool, the results will be of interest both to examinees (medical students and residents) and administrators (medical institutions) and the broader interests of the general public, who are the ultimate recipients of medical services.

The chief theoretical contribution of the dissertation is to better understand personal and situational determinants of job performance evaluations. For example, the results may help us better understand the reasons for the ubiquitous exercise effects observed in the assessment center literature (e.g., Brannick, Michaels, & Baker, 1989; Sackett & Dreher, 1982; Schneider & Schmitt, 1992). The exercise effect is essentially a pattern of high correlations among dimensions within exercises coupled with small correlations among dimension between exercises. The exercise effect is known as “case specificity” in the medical literature. In addition to better understanding the ubiquitous exercise effect, a unique aspect of the current study is the presence of alternate forms of SPs. By examining alternate forms of role plays, we will better be able to disentangle inherent unreliability of physician interactions from unreliability due to differences in scripts applied to standardized patients. The typical OSCE is set up so that different stations tap a partially overlapping set of skills, and so case specificity (exercise factors, e.g., Guiton, Hodgson, Delandshere, & Wilkerson, 2004) are not surprising (c.f., Neidig & Neidig, 1984). This study will be the first to report role play ratings from stations that are designed to be nearly identical because the case is the same in both forms. That is, the current effort will examine what is thought to be the ceiling of reliability and validity for OSCE role plays. This is of theoretical importance because it provides an evaluative standard for evaluations taken in performance simulations.

The organization of the introduction of the dissertation is as follows. First, I will provide a slightly longer description of the OSCE, and review its evolution. Then I will discuss some of the ways in which OSCEs have been assessed psychometrically and what we now know about their reliability and validity. Finally, I will describe the rationale for the current investigation in light of what is currently known about the OSCE and the role-playing standardized patients.

What is an OSCE?

As previously mentioned the OSCE is an examination developed to assess different competencies that are needed to be successful as a medical practitioner (Newble, 2004). It is one of many devices used to test the competency of medical students. Some of the typically assessed competencies include communication skills, history taking, prescription of medication, breaking bad news to patients, and ethics. The choice of competencies is up to the medical college and its intended purpose for the OSCE. The OSCE is typically set up so that the students encounter problems or exercises at a series of stations. In one type of station, students interact with a Standardized Patient (SP) who portrays a scripted ailment; other stations may require students to read radiological films, or to bandage a mannequin. The number of stations can range from just a few to over 20. A student's time at a station typically lasts from 5 to 15 minutes. Each student is allowed the same amount of time at any given station. Each SP (or medical faculty member, or both) rates each medical student on that student's performance at the station. The ratings usually consist of content-specific checklists (e.g., took blood pressure) and also a global rating (e.g., the SP may be asked, "How likely would you be to come back and discuss your concerns with the student again?"). In addition, sometimes an outside observer makes ratings of the medical student on areas such as clinical competency in which a SP might not have expertise.

Recent History of the Clinical Examination

In the 1950s and 1960s, clinical competence was commonly assessed with essays and oral examinations (Newble, 2004), along with short cases and long cases. The following section will briefly touch on how clinical examinations were done shortly before the OSCE was introduced and some of the shortcomings of these examination techniques that led to medical examiners seeking a better way to assess the competence of medical students.

Oral Examinations and Multiple Choice Tests

Before the 1960s, oral examinations were part of the clinical examination that students had to pass in order to become practicing medical doctors. However, in the 1960s, the National Board of Medical Examiners discontinued the use of oral exams after it was discovered that the examiner reliability was unsatisfactory (examiners show large disagreements in judgments of student performance). The solution to the unacceptable reliabilities was the implementation of multiple choice tests. One problem observed with the multiple choice test was that the answer was contained in the choices. Therefore, medical students could rely on recognition to a certain extent rather than on production of the answer (Harden et al. 1975; Mavis, Henry, Ogle, & Hoppe, 1996; Schuwirth & van der Vleuten, 2003).

Another problem with the multiple choice test is that it is well designed to test declarative, but not procedural knowledge. One might be capable of answering questions about surgery, for example, without being able to complete the surgical procedure successfully.

Short and Long Cases

The traditional clinical examination was typically assessed by two examiners who would rate the student's skills on a few (Four in the Wilson, Lever, Harden, Robertson, & MacRitchie, 1969 study) different patients who were divided into short and long cases. Traditionally, these were real patients, untrained for clinical examinations (Wass, Jones, & Van der Vleuten, 2001). Originally, one examiner (A) would rate the medical student for the long case and the other examiner (B) would handle the assessments for the short case (Wilson et al., 1969). This often led to poor correlation between the medical students' scores for the short and longs cases and also a poor correlation with the students' scores on an objective written paper they were also required to write as part of the examination. Often with the short and long cases there was confusion as to what exactly was being tested with each patient and this method also had a chance element for which student got which patient. In addition to these issues, often the examiners had different marking standards which led to low inter-rater reliability (Harden et al.).

Wilson et al. set out to determine if having both examiners rate students on the long and short cases would improve reliability. For the Wilson et al. study, the four patients were divided into one long case that took place over an hour and three short cases for which no time limit was mentioned. In addition to the one-time ratings given during the study, interactions with the patients were video recorded so that other judges (e.g., junior examiners and consultants) could rate the students and there could be a follow-up rating by the original judges of two weeks and two months after the original rating session. The results of this method still showed inconsistency. Examiner A and B's marks on the long case correlated *r* = .78, however for any given candidate, Examiner

A’s rating would often be +/-10 points from Examiner B’s ratings and vice versa. Similar results were found for the short cases with the correlation being a respectable *r* = .84, but examiner A and B’s ratings on an individual candidate being 10 points apart. Examiner A’s marks correlated *r* = .66 between the long case and short case and similarly Examiner B’s marks correlated *r* = .64. When looking at the other 12 judges in addition to the original two examiners, Wilson et al. noted that although 15 students were failed by at least one examiner, none were failed by all examiners. They estimated that out of those 15, nine of them probably deserved to pass. Their solution to this problem was a more objective type of clinical examination advocated by Hubbard, Levit, Schumacher, and Schnabel (1965). Such calls for objective measurement helped spur the development of the OSCE (Harden et al.).

History of the OSCE

Harden, Stevenson, Dowie, and Wilson (1975) first proposed the use of an objective structured clinical examination in lieu of the traditional clinical examination. Reasons for the proposed change to a more objective structured format included taking the luck of the draw out of which student would get what patient and to reduce discrepancies in the rating of the students by introducing checklists for evaluating students based on the specific cases. Also, the inclusion of SPs allowed medical examiners to get a more consistent performance compared to using untrained real patients.

In their original study, Harden et al. examined the correlation between both a traditional clinical examination, referred to as “the clinical” and their proposal for an OSCE. They divided 99 students into 3 equal groups, 66 examined via “the clinical” and the other 33 partaking in the OSCE. All of these students were also given a written multiple choice examination in medicine, surgery, and therapeutics. At the first type of OSCE station an examiner gave the medical student written instructions to either carry out a certain type of procedure or solve a problem that the SP acted out (e.g., determine what led to the SP's shortness of breath). The medical student was given 5 minutes to complete the request and then moved on to the second station. The student was asked either multiple choice or open ended questions about the station he/she just completed. Harden et al. suggested using the multiple choice questions due to ease of marking. Their suggested marking style gave +1 for a correct answer, -1 for an incorrect answer, and 0 for an unanswered question.

Harden et al. suggested using 16 stations for this format. Examiners rated the medical students using checklist with a simple "yes" or "no" being the only ratings for each step of the procedure assessed at a station. This was later revised to allow for a qualified "yes." In addition to the check list, examiners also gave a rating on a five point scale as to the overall proficiency of the student at that particular station. Harden et al. also suggested the use of a SP, but stated that for some types of ailments, colored slides could be used for the students to make a diagnosis. Some of the types of questions suggested by the authors included history taking in a specific area, which of the following were present/absent in the slide you just viewed, which of the following are true about the patient you just examined, etc.

The results achieved by Harden et al. using the OSCE setup compared to the traditional "clinical" were promising. The marks given to the group who took the traditional clinical examination did not significantly correlate with the grades received on a written examination ($\gamma = .17$ and $\gamma = .21$ for groups one and two, respectively). More success was found with the results of the marks for the OSCE and written exam scores as the correlation climbed to $\gamma = .63$, indicating the OSCE may have better criterion-related validity *(Note: γ is the symbol known as the Goodman-Kruskal gamma. It is used for looking at congruence between variables. It is similar to Kendall's τ, except for the denominator).* Some disadvantages of the OSCE mentioned by the authors included increased preparation time by the examiners and the compartmentalization of skills instead of focusing on the whole patient. Harden et al. suggested supplementing the OSCE with the more traditional long case in order to assess a student's ability to look at the whole patient. In the United States and Canada, present day OSCEs are often done as a stand alone assessment without the long case (Wass, Jones, & Van der Vleuten, 2001).

Anecdotal evidence (in the form of medical students describing the SP encounters as "fake") suggests that the OSCE format may have drawbacks as well as advantages. In a recent reconsideration of including real as opposed to standardized patients, Wass et al. (2001) collected data simultaneously on both SP encounters and long cases using real patients. They used two long cases vs. 20 OSCE stations. Each long case lasted 14 minutes where the medical student interviewed the patient in order to gather history and to diagnose the condition and plan for treatment. Each of the OSCE stations was 7 minutes long and the ones utilizing SPs ranged from examining clinical skills (8 stations), practical procedures (4 stations), psychiatry (2 stations), and communication skills (4 stations), with two additional stations on radiology. Using Generalizability Theory, the authors determined that using only one examiner and 8-10 long cases, reliability would surpass .80, thus rivaling reliability findings for the OSCE. Under the format that Wass et al. used, they argued that by using fewer resources (examiners) than with the OSCE, they would be able to get a similar reliability to the 20 station OSCE and that perhaps the long case should not be ruled out as a viable assessment option quite yet. Critics have stated that the use of the long case is too homogenous (Harden et al.) and that is one reason a format like the OSCE was sought after originally. Wass et al. countered with the argument that indeed they are assessing a more homogeneous domain of clinical competence than the OSCE, but any loss of standardization by using real patients and only a few long cases may be compensated for by predicting performance across disciplines as was seen in the Olson (1999) study. The current trend has been to continue on with the OSCE in lieu of the long case.

Since the Harden et al. study, the OSCE has been used for a broad array of medical areas from physical therapy to internal medicine (e.g., Battles, Wilkinson, & Lee, 2004; Hutchinson, Aitken, & Hayes, 2002; Wessel, Williams, Finch, & Gemus, 2003). Pediatrics has seen less use of the OSCE (Carraccio & Englander, 2000), most likely because of the difficulty of finding SPs (child actors) who are believable in pediatric roles. However, there was a trio of articles dealing with exactly this topic published in the early 1980s indicating the versatility of the OSCE (Waterson, Carter & Mitchell, 1980; Watson & Houston, 1982; Smith, Price, & Houston, 1984). Other areas that have more recently adopted an OSCE approach are dentistry (Larsen & Jeppe-Jensen, 2008) and psychiatry (Walters, Osborn, & Raven, 2005). Because of its apparent advantages and NBME endorsement, the format of the OSCE has become widely adopted.

Psychometric Evaluation of the OSCE

As previously noted, the OSCE is typically composed of several stations, each of which is designed to tap one or more clinical skills important for the practice of medicine. Several studies have provided data relevant to the reliability and validity of scores from the OSCE, and these studies may be organized in several ways. For convenience, the studies are divided here into reliability studies and validity studies, even though individual studies often report data that are relevant to both. Data relevant to the reliability of OSCE evaluations are reported first.

It is customary at the end of the encounter at each station for the SP and/or a faculty observer to complete an evaluation form that contains a checklist (e.g., did the student take blood pressure; did the student touch the patient's ankles) and perhaps an overall or summary evaluation of the encounter or of dimensions of interest (e.g., a summary judgment of the quality of history taking). At present, more studies have dealt with the checklist aspect of the OSCE (Park, Chibnall, Blaskiewicz, Furman, Powell, & Mohr, 2004), and fewer have dealt with the global ratings (Amiel, Ungar, Alperin, Baharier, Cohen, & Reis, 2006).

Several different forms or aspects of reliability may be estimated for such evaluations. One may estimate the reliability of checklist scales within stations, which is an internal consistency measure that considers differences in items as a source of error. One may estimate the reliability of checklists and/or the same individual items (such as the overall evaluations) across stations, which provides an internal consistency estimate in which encounters or stations (SPs and their associated scripts) are considered a source of error. One may estimate reliability of sources of the evaluation (SP vs. faculty) on the same station, which is an inter-judge reliability that considers human judgment (but not items or stations) as a source of error. Obviously any given design could consider more than one source of error depending upon how the data were collected, and many studies adopt a generalizability analysis framework in which the impact of multiple facets upon score reliability is examined.

Generally speaking, medical research articles are quite terse, and it is often difficult to determine precisely how the data were collected. For example, in the first article in Table 1 (Brailovsky & Grand'maison, 2000), we know from the article that the data were collected over examinations for several years, and that the examination changed over the years. The first entry in the table concerns the reliability of the history taking scale, which they reported to be .68. But the article failed to report how many stations provided evaluations of history taking (we know that there were 26 stations, and that not all stations provide information on each dimension, but not how many stations for which dimension for which year), nor was it reported how many items comprised the history taking scale, nor how many different judges provided evaluations across stations. The reliability estimates reported for this study in Table 1 are averages taken over 10 different examinations, but there is insufficient information given in the article to allow one to calculate the number of examinees upon which the figures are based. Despite such omissions in reporting, there are important data presented in the literature, and such are summarized in the current paper.

Internal Consistency

Internal consistency estimates have ranged considerably. Acceptable internal consistency reliability is generally given as $\alpha = .70$ for research purposes (Nunnally &

Bernstein, 1994). Table 1 provides a synopsis of the literature. For additional studies see Petrusa (2002). Some studies provided an overall Cronbach's alpha and also provided separate estimates for the different areas that were assessed. Others only give an overall value.

Table 1

Internal Consistency Estimates of Other OSCE Studies

Study	N	OSCE Type	OSCE Stations	Reliability Coeff(s) and Description	Reliability Coeff(s) overall
Brailovsky & Grand'maison (2000)	Spr = 179-262 Fall = 34-75	Licensing exam	26	α = .68 history α = .53 physical exam α = .41 investigation α = .45 diagnostic α = .74 treatment α = .76 communication α = .78 organization	
Wilkinson & Fontaine (2002)	204	Patient education and history taking	11	Global response to single item from SP – how likely to return to doctor; 1 SP/station	α = .65
Wass et al. (2001)	214	Qualifying exam	20	Estimate based on table of ICCs, not actual calc; OSCE includes some non- SP stations	ICC = .65 (est)
Guiton et al. (2004)	421	Communication Skills	7	Average checklist alpha within station; SP rated comm. skills on 7 items	α = .91
Guiton et al. (2004)	421	Communication Skills	7	Alpha of checklists across stations; SP rated comm. skills on 7 items	α = .49
Amiel et al. (2006)	34	Breaking bad news	8	Comm scale w/ 7 common and 3 or 4 unique items per station; SP Likert ratings	α = .81 pretest α = .78 posttest
Park et al. (2004)	286	Psychiatry	9	α = .71 mechanics (binary checklist) α = .85 patient perception questionnaire (Likert scales) α = .73 differential diagnosis α = .67 observation	α = .88
Park et al. (2004)	286	OB/GYN	5	Station scores included both checklists and Likert evals	α = .54
Wessel et al. (2003)	48	Physical Therapy	8	Evaluators were practicing PTs using checklists; 3 written (not SP) stations	α = .48

Table 1 (continued)

Study	N	OSCE Type	OSCE Stations	Reliability Coeff(s) and Description	Reliability Coeff(s) overall
Amiel et al. (1997)	72	Internal medicine	10	α = .56 history (SP) α = .52 physical exam (SP) α = .76 interpersonal skills (SP) α = .66 global rating (physician) α = .31 Oral pres (physician) α = .19 Diff diag (physician) α = .19 Mgmt plan (physician) α = .78 Global oral rating (physician) α = .64 Structured oral exam (physician)	α = .84

Wilkinson, Frampton, Thompson-Fawcett, & Egan (2003)	181 (1997)
	188 (1998)
	205(1999)
	181(2000)
		Qualifying exam	18	Station checklists	α= .83(1997)
					α= .86(1998)
					α= .85(1999)
					α= .88(2000)

Lee, Wilkinson, Battles, & Hynan (2003)	56	History taking and ability to record and interpret data	8	g = .692 (history taking) g = .769 (recording and interpreting data)	g = .672
Schwartz, Witzke, Donnelly, Stratton, & Blue (1998)	N=56 (1993) N=59 (1994) N=51 (1995) N=36(1997)	General Surgery (1993,1994) Head & Neck Exam (1995) Women's Health (1997)	19 (1993) 15 (1994) 17 (1995) 10 (1997)		α= .91(1993) α= .91(1994) α= .78(1995) α= .63(1997)
Newble & Swanson (1983)	429		3-5 (SP only)		α = .31
Vu, Barrows, March, Verhulst, Colliver, & Travis (1992)	405		17 (SP only)		α = .62

Table 1 (continued)

Study	N	OSCE Type	OSCE Stations	Reliability Coeff(s) and Description	Reliability Coeff(s) overall
Shatzer, DaRosa, Colliver, & Barkmeier (1993)	15 (3.5hr version 23 (2hr version)		11 (3.5 hr version) (SP only) 8 (2 hr version) (SP only)	α = .62 (5 min station) α = .82 (10 min) α = .77 (20 min) α = .52 (5 min) α = .60 (10 min)	
Shatzer, Wardrop, Williams, & Hatch (1994)	36		12 (SP only)	α = .77 (5 min station) α = .43 (10 min)	
Stillman, Regan, Swanson, Case, McCahan, Feinblatt, Smith, Williams, & Nelson (1990)	311	Data gathering and interviewing	13 (data gathering) (SP only) 17(Interviewing) (SP only)	α = .68 (data gathering) α = .88 (interviewing)	
Petrusa, Guckian, & Perkowski (1984)	343		10 (SP only)	α = .26-.50	
Matsell, Wolfish, & Hsu (1991)	77		10 (SP only)	α = .12-.69	
Rutala, Witzke, Leko, & Fulginiti (1990)	76		16 (SP only)	α = .94	
Cohen, Rothman, Ross, & Poldre (1991)	36		28 (SP only)	α = .74	
Mann, MacDonald, & Norcini (1990)	89		5 (SP only)	α = .07	
Minion, Donnelly, Quick, Pulito, & Schwartz (2002)	34	Surgery	20		α = .59

Note: The notation (est) means that the number was estimated from data in the article rather than reported in the article.

Numerous authors mention “case effects” or “station effects” such that the evaluative items within a station tend to correlate more highly than do similar items across stations (this is “cases specificity”).

Thus, a fair number of stations are necessary to achieve an acceptable internal consistency estimate across stations. As can be seen in Table 1, acceptable global internal consistency estimates were achieved with as few as 8 stations, but were not guaranteed, even with as many as 26 stations. Thus, internal consistency estimates varied quite a bit across applications, and it appears that simply requiring lots of stations will not be sufficient to guarantee highly reliable scores. Because of the lack of detail in most reports, it is difficult to determine why the reliability estimates vary so highly across papers, although part of the variance could be due to subject matter (e.g., history taking compared to patient comfort).

However, results suggest some interesting possible explanations. First, it appears from the table that global communication skills scores (overall patient perception of the physician’s dealing with the patient) were more reliable than scores indicating more specific skills. Second, it appears that global ratings may be more reliable than checklists, especially when the checklists are targeted to the specific station’s scenario (e.g., differential diagnosis). In other words, the content of the scale may affect the reliability across stations, such that more global and interpersonal characteristics show higher correlations, and more specific and technical characteristics show lower correlations. It is also typically the case that different judges (SPs and/or faculty observers) observed medical students in only one or two stations, so that part of the low correlation among exercises may be due to unreliability attributable to judges. The judge as a source of unreliability is considered next.

Inter-Rater Reliability

As mentioned earlier, one of the problems with earlier medical examinations including oral, essay and actual patient examinations was lack of agreement between judges. By making the OSCE both relatively structured and objective, the designers hoped to minimize disagreements among judges. Typically one would like to see scores for inter-rater reliability above .6 (Nayer, 1993; Schuwirth & van der Vleuten, 2003), but some would argue for at least .8 (Sloan, Donnely, Schwartz, & Strodel, 1995)

Several authors have devised studies examining the degree of agreement between judges. The available studies are summarized in Table 2. Wilkinson, Frampton, Thompson-Fawcett, and Egan (2003) undertook a very elaborate study in order to determine some avenues where improvement might lead to better inter-rater reliability between judges. In this study, OSCEs were set up to be simultaneously run in three different cites in New Zealand. They also included two examiners per station; information about examiner characteristics was gathered (e.g., involvement in the design of the stations, years of experience in both clinical medicine and in conducting examinations). Examiners also observed multiple stations, so that inter-judge reliability could be examined as a function of station. The data were collected in OSCEs for four years, and as time passed, the number of items per checklist was increased.

Using Generalizability Theory, variance components were estimated and the contribution of several variables to inter-rater reliability was assessed. Wilkinson et al.

(2003) found that the degree of involvement in station construction contributed the most to inter-rater reliability. That is, those who designed the station were best able to provide reliable ratings of medical student performance at that station. The researchers' addition of items after each year was intended to make the checklist even more objective to improve reliability, however, there was no relationship between number of items on the checklist and inter-rater reliability (r = .066, p = .62; a partial correlation of number of items and reliability was reported to be -.22, p < .05, but it is unclear what variable(s) were held constant for this analysis). They also found that neither the years of experience in clinical medicine nor years of experience in giving OSCE examinations was related to inter-judge reliability.

As for global ratings, Wilkinson et al. noted that in some other studies (Regehr, MacRae, Reznick, & Szalay, 1998; van Luijk & van der Vleuten, 1992), global ratings may show lower inter-rater reliability between judges at the same station, but their inter-case generalizability (correlation across stations) may actually be higher than those of the checklist. Another study found experienced clinicians may do worse on checklists but better on global ratings than residents and clerks (Hodges, Regehr, McNaughton, Tiberius, & Hanson, 1999). Wilkinson et al. noted that simply increasing the number of items on a checklist may not have the intended result of increasing inter-judge reliability. Instead, it may result in checklists that focus on trivia (see also Newble, Dauphinee, Dawson-Saunders, MacDonald, Mulholland, Page, Swanson, Thomson, & Van Der Vleuten, 1994)

Examiner training has not always been effective in improving inter-rater reliability (Newble, Hoare, & Scheldrake, 1980). Training would seem to be the most direct fix to the problem and a different type of training may lead to different results. Table 2 presents some previous studies that looked at inter-rater reliability using both global and individual area ratings. Typically in an OSCE, it is examiners doing the rating and the SP serving as part of the assessment.

Table 2

Inter-Rater Reliability of OSCEs From Selected Studies

Study	N	Type of OSCE	OSCE Stations	What was rated?	Reliability estimates	Number of Judges
Sibbald & Regehr (2003)	108	Pharmacy	12 (each student had 4 cases and there were 3 cohorts)	Global rating; 5 Likert items, (empathy verbal skills, etc., overall)	Pro SP α=.44 Student SP α=.56	2 (SP and expert examiner)
Wilkinson, Frampton, Thompson-Fawcett, & Egan (2003)	181 (1997) 188 (1998) 205 (1999) 181 (2000)	Qualifying exam	18	History Examination Investigation Management Patient Education	r=.76(1997) r=.78(1998) r=.80(1999) r=.76(2000)	2 per station; mean inter-judge corr across years using checklist was r = .78
Wass et al. 2001		Long case revisited		Six global ratings on Likert scale summed and expressed as percent of possible total. No rater training of expert judges (physicians)	Generalizability analysis of long case; 2 examiners for 1 long case,	ICC = .41
Walters 2005		Psychiatry OSCE			Generalizability coefficient considering examiner, student, station. Reported ICC not just for examiner	Range across circuits ICC = .55 to .68; however, examiner variance small compared to student and station

Table 2 (continued)

Study	N	Type of OSCE	OSCE Stations	What was rated?	Reliability estimates	Number of Judges
Quest, Ander, & Radcliff, 2006		Death disclosure	ACS, affective competency score	Global rating of communication competency	Computed student, faculty, and SP correlations; single scenario with SP	SP vs. Faculty rating r = .47 student vs. SP rating r = -.04 Student vs. faculty r = .00

What, if anything, can we conclude about the inter-judge reliability of the OSCE? Newble (2004) reported that suitably trained judges can provide reliable ratings, and that "…global ratings, within the framework of structured tasks and used by informed or trained assessors, may be as reliable or even more reliable than checklists" (p. 201). Newble also advocated matching the sort of evaluations scheme (global rating versus checklist) to the sort of dimension being rated. He noted that more technical (presumably what we would consider procedural knowledge) aspects are well suited to checklists, whereas other skills that involve interpersonal interaction, such as communication or other process skills, are better suited to global ratings. Based on the data reviewed here, inter-judge reliability on checklists appears adequate (approaching .80). However, the reliability for scores from global ratings across examiners and standardized patients appears lower, and the agreement of expert examiners for the performance of students on a single case appears unacceptably poor.

Common sense suggests that checklists for procedural items (e.g., did the physician take the patient's blood pressure) will show good reliability, provided that they are properly matched to the procedures required by the task and easily visible to the examiner. However, judgments that are more qualitative, such as competence in communication, are likely to show poorer inter-judge reliability. There may be a tradeoff between reliability within and between stations when comparing global evaluations and checklists, but this may not be a function so much of the format of the item as of the nature of the dimension to be measured.

Content Validity

Newble (2004) listed the following steps to support content validity: First, one must identify the different problems and conditions in which the medical student needs to have competence. For the second step, one must define the tasks in response to the problems or conditions that were set up in step one. The example that Newble (2004) gives for a condition is for a SP complaining of “chest pain.” To score adequately at this station, it would be necessary for the medical student to accomplish such tasks as taking a medical history, requesting and then interpreting an ECG, etc. For the third step, Newble (2004) recommends making a blueprint. This consists of a two-dimensional matrix with generic competencies to be tested represented on one axis (e.g., history taking) and the problems or conditions (e.g., chest pain) where the competencies will need to be demonstrated on the other axis. A good example of such a blueprint can be seen in Tombleson, Fox, and Dacre (2000).

Criterion-Related Validity

Shibald and Reger (2003) looked at criterion related validity by using professional SPs and also 1st year medical students trained as SPs. Two different criterion measures were used: scores on a written exam, and clinical marks. The predictors were global ratings of empathy, coherence (organization and focus), verbal skills, nonverbal skills, and overall impression (knowledge and skills integration). In addition to having the SPs rate the students; they were also rated by expert observers. When professional SPs played the role of patients, concurrent validity was *r* = .44 and *r* = .26, respectively for expert observer and SP rating scores with the course written exam. Predictive validity correlations were down slightly when the ratings of the expert observers and SPs were compared to the clinical mark at *r* = .23 and *r* = .14, respectively. The concurrent validity dropped slightly when student patients were used as SPs compared to the professional SPs. The concurrent validities in this case were *r* = .19 and *r* = .16 for the medical students when rated by the expert observers and the student SPs, respectively. Predictive validity increased slightly with correlations of *r* = .30 for the expert observers and *r* = .24 for the student SPs when compared to the final clinical mark.

Wessel et al. (2003) tried to establish predictive validity of a physical therapy OSCE by comparing ratings on the OSCE to a previously validated instrument. Their OSCE had eight stations, of which five were role plays and three were written. The instrument they chose for a criterion was the Physical Therapist Clinical Performance Instrument (CPI) which consisted of 24 items. Only six items from the CPI were chosen for this validation because these six items matched up with the skills that the OSCE was supposed to assess. The items chosen by the authors were (1) safety, (6) communication, (11) physical therapy examination, (12) interpretation of findings, (14) performance of interventions, and (15) the education of others. Unfortunately, along with low internal consistency ($\alpha = .48$) for the OSCE, the OSCE also failed to predict clinical performance as assessed by the CPI (the overall correlation between the OSCE and the sum of CPI scores was -.13). Reasons for the poor results mentioned by the authors include the CPI being more of a global assessment than the OSCE and this OSCE being the first such exam for the students involved in the study.

Construct Validity

Implicit in the enumeration of competencies and ratings of multiple dimensions of performance is the notion that the dimensions are sufficiently distinct, at least conceptually, so that ratings of different dimensions are useful for assessment, feedback, or some other administrative purpose. In other words, one would expect to see distinct dimensions to yield distinct ratings, or discriminant validity. Although most OSCEs collect data around multiple dimensions (e.g., history taking, diagnosis, communication, etc.), to date there are few studies that examine the construct (factorial) validity of the evaluations thus gathered. Assuming valid assessment, one would expect that correlations of similar dimension across cases or stations would correlate more highly than would different dimension within cases. For example, one would expect ratings of history taking across abdominal pain and trauma cases would correlate more highly than ratings of history taking and diagnosis within the abdominal case and within the trauma case. The few studies that have examined ratings for the desired pattern of results have found the pattern of results opposite to that desired. Factor analysis of ratings yields factors that correspond to cases rather than to dimensions (e.g., Guiton et al., 2004). Brailovsky and Grand'maison (2000) looked at construct validity of an OSCE using the multitrait-multimethod (MTMM) approach. For this study the authors utilized three assessment methods; the Quebec SP-based exam (OSCE) and two instruments used in the certification examination of the College of Family Physicians of Canada (CFPC). One of these exams was the multiple choice questions test, which was still in use at the time and the short answer management problems (SAMPs) of the exam. Brailovsky and Grand'maison stratified the content of all three tests into problem definition and management, two attributes of clinical competence needed to succeed as a physician. The results of this study showed that the method of measurement was more important than the attribute being measured. That is, the correlation of different traits measured by the same method was greater than the correlation of the same traits with different methods. No specific numbers were provide for this MTMM, but the conclusions the authors drew from this study were that clinical competence is a very complex and comprehensive construct and that multiple methods that are deemed different, yet complementary, should be used to assess clinical competence.

The Current Study

Case specificity is a very common finding in the OSCE literature. Operational OSCEs, however, do not typically include alternate forms of a case; if there are alternate forms, they are not administered to the same students (Swanson, Clauser, & Case, 1999). Thus the tacit assumption that medical student behavior is consistent within the same case is never tested. The level of medical student consistency of behavior within situations sets a ceiling on the reliability and validity of OSCE ratings between situations. The current study is unique in that it contains alternate forms of two cases. These alternate forms allow one to examine what is essentially the ceiling for reliability and validity of ratings taken in the OSCE. The psychometrics of the results across alternate forms can then be used as a sort of baseline or yardstick by which to evaluate the reliability and validity of the rest of the stations.

Data for the current study were obtained in a Comprehensive Clinical Performance Exam (CPX) given by a medical school that utilized the OSCE format. In this CPX, 12 of the stations were role plays using SPs and we looked at two stations from the fall and two from the spring which will be elaborated on later. SPs also served as the examiners and rated the medical students on interpersonal skills like confidence and comfort, and also completed checklists detailing student performance on the specific requirements of the station (e.g., history taking, physical exam, etc.).

Perhaps unique to this particular examination, there are two pairs of stations in which the case being examined is essentially repeated. For the first case, the patient is a female suffering from abdominal pain and the second case features a male patient with pneumothorax. Because a set of medical students participated in all four exercises, the correlations of all dimensions can be computed across exercises, yielding a multitrait multimethod (MTMM) matrix where aspects such as communication and history taking serve as traits and stations serve as methods. The cases, as presented, were not exactly alternate forms in that some of the items differed and this will be elaborated upon later. However, we will analyze the exam both as it was given and also using just the common items between the cases.

Additionally, Generalizabilty Theory will be used (G-Theory) for this analysis in order to determine the number of examiners and stations needed to reach a threshold reliability of .80 deemed necessary for a high stakes examination (Swanson, 1987; Crossley, Davies, Humphris & Jolly, 2002) for the assessment of communication skills, history taking, and clinical exam skills. G-Theory was introduced by Cronbach (1963, 1972) in response to limitations present in the Classical Test Theory. Classical Test

Theory states that you have an observed score, X, a true score, T, and error, E that is arranged in the following formula: X=T+E. This essentially means that the observed score is a reflection of someone's true score and an error term that makes the observed score deviate from the true score. The problem with this model is that while it does account for error, there is no specification as to where the error is coming from. It may be suitable to use the Classical Test model for carefully equated parallel forms, but when alternate types of tests are used the Classical Test model becomes to restrictive (Matt, 2001). A full explanation of G-Theory is beyond the scope of this paper, but essentially, G-Theory will allow us to partition the error variance into different facets including rater, station, and ratee.

Chapter Two

Method

Participants

Participants were medical students from a large university in the southeastern United States. At the present time, using archival data, there was information for about 135 participants who were assessed using the alternate forms of the SPs representing similar ailments (there are approximately 120 students per year in the medical school and two years of data; numbers of student participants included in this study were somewhat smaller due to missing data and only certain students from each year seeing all 4 cases of interest to this particular study). The gender break down was approximately 52% female and 48% male. All of the students who participated were 3rd year medical students. Neither SP nor medical student identities were linked to records analyzed in this project. Each participant and SP was given a unique study identification number that cannot be linked to the person's identity and IRB approval was secured.

Procedure

Test Development. Four role-plays were developed by the medical school for the CPX. The four role names were Rachel Brown (RB), Samantha Browning (SB), John Sexton (JS), and John Long (JL). The two female roles portrayed the same case, which involved abdominal pain (appendicitis). The male role was slightly different. They are essentially the same case, however one was played by a mannequin voiced by an SP who viewed the encounter through a two-way mirror and the other was an actual SP. The diagnosis for this case was pneumothorax. The cases were developed to satisfy requirements set by the National Board of Medical Examiners (NBME) for certification. Each case has a detailed script for the SP to use in portraying the patient. The script contains information concerning history, symptoms, and associated materials such as cards that indicate the results of invasive procedures not carried out during the encounter with a simulated patient (e.g., there may be a card indicating the results of a rectal examination). The case also includes a medical chart given to the medical student participant, indicating what is known about the patient through intake, just as would ordinarily be available when a doctor sees a patient.

The RB case consisted of seven communication items, 15 history taking items, and 16 clinical examination items. The SB case was similar, consisting of 12 communication items, 15 history taking items, and 16 clinical examination items. For history taking and clinical examination, RB and SB were exactly the same and the communication dimensions were very similar. The JL case had 12 communication items, 11 history taking items, and four clinical examination items. JS had seven communication items, 12 history taking items, and nine clinical examination items. These two cases shared five history taking items and two clinical examination items. The first set of data analysis will deal with the four cases as they were presented to the medical students. The second set of analysis will deal with only the common items between RB and SB and the common items between JL and JS.

Test Administration. Participants completed the CPX from July 2006 to April 2008. More specifically, the mini CPX which contains the roles of Rachel Brown (RB) and John Sexton (JS) was administered in July, September, November, January, March, and May of 2007 and 2008. The items for the communication dimension for these two cases were exactly the same. The comprehensive CPX containing the roles of Samantha Browning (SB) and John Long (JL) took place from April to May of 2007 and 2008. As with RB and JS, the items for the communication dimension for these two cases, SB and JL, were exactly the same. All role-plays took place in a clinic designed for medical simulations. All role-plays with medical students and standardized patients were captured by audio/video recording devices and stored electronically. Each encounter lasted 11 minutes from beginning to end.

Test Scoring. For the scenarios considered here, SPs provided evaluations of three dimensions (Communication, History Taking, and Physical Examination), and medical school faculty or administrators provided evaluations for the fourth dimension (Critical Thinking) which was not evaluated. After each encounter, SPs completed a series of evaluations on a standardized form. Several questions concerned Communication. These were assessed using Likert-type items rated on a scale of 1-5 regarding the quality of the interaction between the doctor and patient (e.g., the doctor explained things well). The SPs also completed checklists that form the basis of evaluations of History Taking (e.g., did the doctor ask when the pain started?) and Physical Examination (e.g., did the doctor touch my ankles?).

After the encounter, the medical student answers a series of open-ended questions regarding the case (e.g., what is the primary diagnosis? What tests would you order?). A faculty member or trained administrator reviews the student’s answers to the questions and grades them based on a rubric designed for the case. The scores on each dimension for a station are weighted and combined with scores on the other stations to arrive at an overall score for the CPX, but overall CPX scores are not used in the current project.

Raters. Standardized patients paid for their work in the CPX provided one set of ratings for each participant. These ratings were also stored electronically by the clinic. The training for all paid SPs consists of the following. Regular training sessions take place during the year. SPs are accepted based upon the demographic needs of the case. Typically SPs play only one case (except for the four mentioned, which have different names, but are otherwise identical). SPs learn their job requirements via lecture and slides and then must play the role for the physicians before they are certified to be a SP. During the actual examination, SPs have access to their respective scripts until a student enters the room. This allows them to be as consistent as possible. For this study, the SP was played by as many as three different people per case. We treated this as a hidden facet, that is, the analysis was completed as if each role was only played by one person.

In addition to the ratings we received from the SPs and physicians, each video recording of the role-play was viewed by up to five additional raters. One of the raters was a graduate student in industrial/organizational psychology; the others were either graduate or undergraduate students. All of the additional raters obtained the same training as that given to paid SPs and used the same evaluation forms used by the paid SPs. Additionally, the raters were trained on rating the communication portion of the evaluation form using a Behaviorally Anchored Rating Scale (BARS). The BARS was created by interviewing several SPs in order to determine what behavior needed to be exhibited by the medical student in order to receive a certain rating. For example, in order to score 3 out of 5 on the introduction, the medical student had to make eye contact with the SP. In order to receive a 4, they additionally had to shake the SP's hand. Behavioral referents were created for each communication item. The additional student raters were utilized as a standard of comparison for the SP's ratings. The literature (Swanson et al., 1999) has shown that typically variance accounted for by rater is relatively small, but for this project we would like to be as stringent as possible. Each rater saw between 258 and 282 videos. The data collection design for raters is shown in Appendix A. One SP and two random student raters out of five provided the scores and then the average of the student raters' scores for each scenario was taken. This will be broken down more specifically in the results section.

Reliability and Validity

We computed a MTMM matrix where the dimensions on which the medical students were being assessed represent traits and the stations represent methods. Note that the structure of the matrix is rather unusual. Two segments of the matrix (the intercorrelations of the dimensions for RB and SB is one such; the other is the intercorrelation of JS and JL) show what are essentially alternate roles of the same case. There are two validity diagonals contained here, one for each pair of within-case roles.

The matrix so constructed was interpreted according to the Campbell-Fiske (1959) criteria for establishing convergent and discriminant validity. Additionally, we computed mean correlations within each method (heterotrait-monomethod correlations) for each of the four stations. Means of the validity diagonal correlations within cases were computed, that is, means were computed for alternate roles within cases. We also computed means of the heterotrait-heteromethod correlations within the alternate roles. Finally means of the validity diagonals and of the heterotrait-heteromethod correlations were computed between cases. The magnitude of the mean correlations indicated the influence of context on the measures.

Within each pair of matched stations (within cases), the validity diagonal was essentially an alternate forms reliability estimate. Because the two exercises employ the same case, we expected to see the maximum convergence of measures of history taking, physical exam skills, and so forth. The two different cases (abdominal pain vs. pneumothorax) provide more conventional evidence of convergent and discriminant validity. If the evaluations across cases follow their customary pattern, we expect to see correlations indicative of exercise or case factors. One unique feature of this study was the ability to compare the convergent and discriminant validity when cases were the same and when cases were different. Additionally, an exploratory factor analysis (EFA) was computed to determine whether factors corresponding to methods, traits, or both would emerge from the correlation matrix of all ratings of student performance. We used the maximum likelihood extraction VARIMAX rotation.

The mean correlation between judges within roles was computed for each dimension of the SP form. It was predicted the correlation between judges would be greater for the checklist dimensions (History Taking and Clinical Examination) than for the Communication dimension. The mean correlation across roles between judges was also computed. It was predicted that the correlation between judges across roles would be greater for Communication than for History Taking and Clinical Examination.

We also used generalizability theory (G-Theory) for this analysis in order to determine the optimal number of examiners or raters and stations needed to reach a threshold reliability of .80, mentioned earlier, for the assessment of communication skills, history taking, and clinical exam skills. Using the information we find in our Generalizabilty Study (G-Study), we performed a Decision Study (D-Study), which is where we take what we have learned from the data of a G-Study and apply it. An example of this is determining how many raters and stations we would need to meet the .80 reliability deemed necessary for high stakes examination. The model for using Generalizability Theory in the current study consisted of multiple facets, rater (6 levels) and station (4 levels). The dependent variable was the dimension score, that is, the sum of items from the SP evaluation or the score assigned by the scoring rubric. All of the facets were considered random, as each of these was considered to be sampled from a larger universe of raters, cases, and roles. For the set of four stations, variance components and generalizability (reliability) coefficients were reported for each dimension, indicating the estimated reliability of the current four stations, and also of feasible combinations of raters, cases, and roles that should result in overall reliability of .80. The results were computed assuming relative error variance.

Chapter Three

Results

The results are arranged into two main subsections. The first deals with reliability; the second deals with validity.

Reliability of Ratings

Abbreviations. Throughout the results and discussion sections the following abbreviations will be used: CM - communication dimension, HX - history taking dimension, and PX - physical examination dimension.

Overview of Analyses. The reliability analyses proceed from the simple to the complex. First, descriptive statistics for the judges, dimensions, and scenarios are presented. Second, reliability estimates are presented for the scores as they are currently organized, that is, reliability estimates are presented for dimension scores for each exercise. Finally, generalizability theory is used to compute variance components overall and for each dimension for raters, cases, and students (ratees) and their interactions. The variance components are then combined to compute estimates of the number of judges and cases needed to achieve a reliability coefficient of .80 for a hypothetical examination similar to the one studied here.

Additionally, the results are first presented on the entire scales reported for each case as collected for the CPX. The items used during the CPX differed somewhat depending on the specific role, so that some of the items in the history scale, for example, are identical across roles, but others are not. Therefore, a second analysis is also presented in which only identical items across alternate roles are included. The deletion of unique items affects the appendicitis cases (RB and SB) less as the only differences occurred in the communication dimension where they shared six (out of seven) items. All 15 items for history taking and all 16 items for physical examination were shared. For JS and JL, six communication items were identical (JL had 12; JS had 7). However, for history taking, there were only five identical items and for the physical examination portion, there were only two.

Descriptive Statistics

First the descriptive statistics of the raw scores overall and by case are presented in Tables 3 and 4. Information is presented separately for student raters and for the SP. The number of ratees varies slightly by case due to technical problems with some of the videos.

Table 3

Descriptive Statistics of Raw Scores for the Four Cases

Case	n	Dimension	Total Score	Min	Max	Mean	SD	Source
RB	132	Overall	66	31	57	44.23	5.33	Rater
				36	66	55.49	6.18	SP
		CM	35	19	35	25.76	2.8	Rater
				22	35	33.19	3.0	SP
		HX	15	2	14	8.43	2.3	Rater
				5	15	10.45	2.5	SP
		PX	16	2	16	10.04	2.9	Rater
				4	16	11.85	3.0	SP

Table 3 (continued)

Case	n	Dimension	Total Score	Min	Max	Mean	SD	Source
SB	133	Overall	91	46	87	63.61	7.1	Rater
				62	91	78.87	6.8	SP
		CM	60	31	60	44.69	5.4	Rater
				42	60	55.89	4.8	SP
		HX	15	4	15	8.90	2.4	Rater
				4	15	11.38	2.3	SP
		PX	16	0	16	10.01	2.6	Rater
				3	16	11.60	2.5	SP
JL	134	Overall	75	30	63	48.04	6.7	Rater
				40	75	65.13	8.6	SP
		CM	60	24	57	40.27	5.8	Rater
				32	60	53.07	7.5	SP
		HX	11	2	10	6.82	1.8	Rater
				4	11	8.90	1.9	SP
		PX	4	0	4	0.94	1.1	Rater
				0	4	3.16	0.90	SP
JS	133	Overall	56	19	50	35.62	5.7	Rater
				18	56	46.12	6.9	SP
		CM	35	13	33	23.93	3.5	Rater
				13	35	31.20	5.0	SP
		HX	12	1	12	6.59	2.3	Rater
				1	12	8.11	2.5	SP
		PX	9	0	9	5.10	1.9	Rater
				2	9	6.81	1.6	SP

Table 4

Common Item Descriptive Statistics of Raw Scores for the Four Cases

Case	n	Dimension	Total Score	Min	Max	Mean	SD	Source
RB	132	Overall	61	24	53	40.60	4.84	Rater
				32	61	50.82	5.86	SP
		CM	30	15	30	22.12	2.09	Rater
				18	30	28.53	2.56	SP
		HX	15	2	14	8.41	2.19	Rater
				5	15	10.45	2.54	SP
		PX	16	1	16	10.07	2.80	Rater
				4	16	11.84	3.01	SP
SB	133	Overall	61	28	57	41.55	4.77	Rater
				37	61	51.02	4.95	SP
		CM	30	15	30	22.73	2.23	Rater
				22	30	28.14	2.47	SP
		HX	15	0	15	8.86	2.31	Rater
				4	15	11.33	2.33	SP
		PX	16	0	16	9.96	2.51	Rater
				3	16	11.55	2.51	SP
JL	134	Overall	37	14	32	24.53	2.92	Rater
				20	37	33.25	4.24	SP
		CM	30	10	27	20.10	2.37	Rater
				17	30	27.05	3.74	SP
		HX	5	1	5	3.78	0.94	Rater
				1	5	4.31	0.93	SP
		PX	2	0	2	0.65	0.83	Rater

Table 4 (continued)

Case	n	Dimension	Total Score	Min	Max	Mean	SD	Source
JL	134	PX	2	0	2	1.90	0.33	SP
JS	133	Overall	37	16	34	24.57	3.14	Rater
				12	37	32.20	4.81	SP
		CM	30	13	28	20.57	2.49	Rater
				11	30	26.79	4.28	SP
		HX	5	0	5	3.12	0.95	Rater
				1	5	3.91	1.05	SP
		PX	2	0	2	0.89	0.85	Rater
				0	2	1.50	0.80	SP

In table 5, the raw scores of the original exam seen in Table 3 were converted to percentages. Due to the averages of the ratings appearing to be different between the raters and SP, a two sample t-test assuming equal variances was performed for each dimension by case. All of the average ratings on total score and on each dimension were significantly different.

Table 5

Descriptive Statistics of Raw Scores Converted to Percentage with Results of Two-Sample t-test

Case	Dimension	Average Rater % Score	Average Rater SD	SP % Score	SP SD	t-value	sig?
RB	Cm	0.737	0.004	0.948	0.008	-22.1	Yes
	Hx	0.563	0.02	0.697	0.029	-6.93	Yes
	Px	0.626	0.03	0.740	0.035	-5.06	Yes
	Total	0.671	0.075	0.840	0.094	-16.1	Yes
SB	Cm	0.746	0.005	0.933	0.006	-19.6	Yes
	Hx	0.595	0.0240	0.755	0.024	-8.33	Yes
	Px	0.620	0.024	0.722	0.025	-5.23	Yes
	Total	0.699	0.07	0.866	0.075	-18.7	Yes
JL	Cm	0.672	0.006	0.889	0.014	-17.5	Yes
	Hx	0.621	0.024	0.813	0.03	-9.4	Yes
	Px	0.237	0.06	0.794	0.044	-19.7	Yes
	Total	0.640	0.074	0.873	0.11	-20	Yes
JS	Cm	0.685	0.007	0.893	0.02	-14.6	Yes
	Hx	0.548	0.033	0.678	0.044	-5.36	Yes
	Px	0.565	0.037	0.757	0.034	-8.22	Yes
	Total	0.636	0.088	0.825	0.122	-14.3	Yes

Note: t-crit for p<.05= 1.969

ICC values

ICCs were based on the assumption of random raters. For these calculations there were 130 students with complete data. Results are presented in three tables. In each table, results were calculated first using both psychology raters and the SP, and second, using psychology raters only. Because of the large difference in means between the raters and the SPs, the reliability estimates are lower when the raters and SPs are combined. The first of the three tables (Table 6) shows results for all scales, items, and judges; Table 7 shows results for the common items, and Table 8 shows estimates for a single, random judge. Communication ICCs were much stronger for the raters trained with the BARS than when SPs were included. The estimates in Table 8 correspond most closely to the way in which the current CPX is administered (there is only a single judge for each case; the judge may not be the same person for all examinees)

Table 8 shows (for the psychology rater data) that the reliability estimates for communication are generally unacceptably low for interpretation at the case level, the reliability estimates for history taking tend to be a bit low but approaching acceptable levels (in the .70s), and the estimates for physical exam are quite variable, with some exceeding .80, but others showing estimates less than .60.

Table 6

ICC of Ratings by Dimension

Case	Dimension	*r* with 3 random raters (2 student raters and SP)	*r* with 2 random student raters
RB	CM	.25	.52
RB	HX	.78	.88
RB	PX	.82	.88
SB	CM	.29	.52
SB	HX	.73	.93
SB	PX	.83	.91
JL	CM	.34	.61
JL	HX	.58	.85
JL	PX	.31	.80
JS	CM	.30	.58
JS	HX	.82	.82
JS	PX	.64	.73

Table 7

Common Item ICC of Ratings by Dimension

Case	Dimension	*r* with 3 random raters (2 student raters and SP)	*r* with 2 random student raters
RB	CM	.25	.48
RB	HX	.78	.88
RB	PX	.81	.85
SB	CM	.29	.51
SB	HX	.75	.92
SB	PX	.82	.91
JL	CM	.24	.47
JL	HX	.70	.79
JL	PX	.40	.80
JS	CM	.26	.57
JS	HX	.74	.77
JS	PX	.65	.84

Table 8

ICC of Ratings Using One Random Judge

Case	Dimension	*r* including SP and student rater data	*r* using only student rater data
RB	CM	.10	.36
RB	HX	.54	.78
RB	PX	.60	.79
SB	CM	.12	.35
SB	HX	.48	.86
SB	PX	.62	.84
JL	CM	.14	.44
JL	HX	.31	.74
JL	PX	.13	.67
JS	CM	.12	.41
JS	HX	.60	.70
JS	PX	.37	.58

G-studies/D-studies

Interpretation of individual case or exercise performance may be desirable for feedback, and the reliability estimates in Tables 6 through 8 would apply in such a situation. However, it is generally the case that decisions are based on exams composed of multiple cases rather than individual exercises. Usually the cases are evaluated by different judges or raters. The reliability of such examinations can be estimated through generalizability theory. Generalizability theory allows the estimation of variance components of the facets of the design (here cases and raters). The variance components can be combined to estimate the reliability of different hypothetical exams. Both sorts of computations are presented in this section.

For this portion of the analyses, we computed a G-study and a D-study on the overall data. Then we computed G-studies and D-studies on each of the dimensions, namely communication, history taking, and clinical examination. Negative variance components are reported for the G-study for verification purposes, but for the D-studies, all negative variance components were set to zero.

Overall G-study

Generalizability theory is typically computed using random-effects analysis of variance. There is a literature on methods of analysis for different data collection designs. The design for data collection need not match exactly the design for data analysis, but the analysis can become very difficult, and often requires a statistician to ensure that the proper variance components are estimated correctly. When all the facets are completely crossed, however, the analysis is simplified. Therefore, a subset of 20 examinees was randomly drawn and rated by all five raters, so that a completely crossed rating design was available for the generalizability analysis (see Appendix A for a schematic of the full rating design).

The variance components for the overall score using all five raters and the SP are shown in Table 9. The largest proportion of variance came from the rater at almost 35%, with relatively little of the variance coming from the medical student at just under 6%. Results for the common item data are in Table 10, with similar results. In Table 11, when the SP ratings are removed, the variance accounted for by rater drops to around 9 percent and the amount that the medical students contribute rises to just over 10 percent. In Table 12, looking at the common item data for the overall score, variance accounted for by the medical student, rater, and case drops slightly while variance accounted for by medical student x case increases.

Table 9

Contribution of Each Source of Variance to the Overall Score Using Five Student Raters and SP

Source	Variance Component	Relative Contribution
Medical Student	0.00079	5.72%
Rater	0.004779	34.62%
Case	0.001319	9.56%
Medical Student x Rater	0.000373	2.70%
Medical Student x Case	0.002759	19.98%
Rater x Case	0.000454	3.29%
Medical Student x Rater x Case	0.003333	24.14%

Table 10

Common Item Contribution of Each Source of Variance to the Overall Score Using Five Student Raters and SP

Source	Variance Component	Relative Contribution
Medical Student	0.000522	3.40%
Rater	0.005362	34.87%
Case	0.000602	3.91%
Medical Student x Rater	0.000404	2.62%
Medical Student x Case	0.004096	26.63%
Rater x Case	0.000435	2.83%
Medical Student x Rater x Case	0.003957	25.73%

Table 11

Contribution of Each Source of Variance to the Overall Score Using Five Student Raters

Source	Variance Component	Relative Contribution
Medical Student	0.000916	10.27%
Rater	0.000809	9.06%
Case	0.001376	15.43%
Medical Student x Rater	0.00039	4.37%
Medical Student x Case	0.002838	31.81%
Rater x Case	0.00028	3.13%
Medical Student x Rater x Case	0.002313	25.92%

Table 12

Common Item Contribution of Each Source of Variance to the Overall Score Using Five Student Raters

Source	Variance Component	Relative Contribution
Medical Student	0.000658	6.80%
Rater	0.000717	7.41%
Case	0.000764	7.89%
Medical Student x Rater	0.000466	4.81%
Medical Student x Case	0.004115	42.50%
Rater x Case	0.000344	3.55%
Medical Student x Rater x Case	0.002617	27.03%

Overall D-Study

The results for the overall D-study can be seen in Table 13. To achieve a G-coefficient of .80, which is the generally accepted minimum value for a high stakes examination, for the overall exam (using all 3 dimensions) using one rater, 34 stations are needed. If using two raters, then only 24 stations are needed and going to four raters lowers the number of stations to 19. For the common item data (Table 14) the number of cases needed with four raters increases to 40. The pattern was similar when using one and two raters with one rater requiring 65 cases and two raters needing 49 cases to reach the .80 threshold. When using the results from only my raters for the overall score, as seen in Table 15, these numbers change to 24, 19, and 16 stations for one, two, and four raters, respectively. In Table 16, the number of raters needed when using the common item data increases to 30, 35, and 44 cases when using four, two, and one rater, respectively. Numbers of cases varies across tables because the number of cases were chosen so that the reliability estimate would exceed .80 for each number of judges.

Table 13

D-study for Overall Score Using Five Student Raters and SP

	Number of Cases			
Number of Raters	12	19	24	34
1	0.591	0.696	0.743	0.804
2	0.671	0.763	0.803	0.852
4	0.719	0.802	0.836	0.879

Table 14

Common Item D-study for Overall Score Using Five Student Raters and SP

	Number of Cases			
Number of Raters	12	40	49	65
1	0.424	.711	.751	.800
2	0.499	.769	.803	.844
4	0.546	.801	.831	.867

Table 15

D-study for Overall Score Using Five Student Raters

	Number of Cases			
Number of Raters	12	16	19	24
1	0.669	0.730	0.762	0.802
2	0.727	0.780	0.808	0.842
4	0.759	0.808	0.833	0.863

Table 16

Common Item D-study for Overall Score Using Five Student Raters

	Number of Cases			
Number of Raters	12	30	35	44
1	.527	.736	.765	.804
2	.585	.779	.805	.838
4	.619	.803	.826	.856

Communications G-study

Table 17 indicates the results of the G-study for the communications dimension. This was by far the most subjective dimension and was rated using a Likert-type scale where each item was rated on a scale from 1-5. With my raters and the SP rater, the variance accounted for by rater was 43.62%. Results were similar for common item data shown in Table 18. The percentage of variance accounted for by rater drops to 13.01% when the SP ratings are removed and only the BARS trained raters are used (Table 19). This drop is also evident in the common item communication data seen in Table 20. However, in all four models only around 5% of the variance for communication is contributed by the medical student indicating little true score variance.

Table 17

G-study of Communication Dimensions Using Ratings from Five Student Raters and SP

Source	Variance Component	Relative Contribution
Medical Student	0.000625	4.06%
Rater	0.006711	43.62%
Case	0.000863	5.61%
Medical Student x Rater	0.000444	2.89%
Medical Student x Case	0.001712	11.13%
Rater x Case	0.000367	2.38%
Medical Student x Rater x Case	0.004663	30.31%

Table 18

Common Item G-study of Communication Dimensions Using Ratings from Five Student Raters and SP

Source	Variance Component	Relative Contribution
Medical Student	0.000733	4.33%
Rater	0.006672	39.47%
Case	0.001431	8.47%
Medical Student x Rater	0.000662	3.92%
Medical Student x Case	0.002018	11.94%
Rater x Case	0.000472	2.79%
Medical Student x Rater x Case	0.004917	29.09%

Table 19

G-study of Communication Dimensions Using Ratings from Five Student Raters

Source	Variance Component	Relative Contribution
Medical Student	0.000584	6.76%
Rater	0.001124	13.01%
Case	0.000962	11.14%
Medical Student x Rater	0.000534	6.19%
Medical Student x Case	0.001764	20.42%
Rater x Case	0.000335	3.87%
Medical Student x Rater x Case	0.003337	38.62%

Table 20

Common Item G-study of Communication Dimensions Using Ratings from Five Student Raters

Source	Variance Component	Relative Contribution
Medical Student	0.000727	7.06%
Rater	0.001084	10.52%
Case	0.001706	16.56%
Medical Student x Rater	0.000783	7.60%
Medical Student x Case	0.002126	20.64%
Rater x Case	0.000361	3.50%
Medical Student x Rater x Case	0.003514	34.12%

Communications D-study

These results are laid out in Tables 21 thru 24. To achieve a G-coefficient of .80 for the communication dimension, 44 stations are needed for one rater. When using two raters, the number of stations becomes 28, and using four lowers the station number to 20. These numbers change slightly for common item data seen in Table 22. Here, using one rater, 41 stations are needed and using two raters drops the number of stations to 26. Using four raters allows the use of 19 stations to reach the desired reliability level. When using only the student raters who received BARS training (Table 23), those numbers change to 41 stations for one rater, 26 for two, and drops to 19 stations for four raters. These numbers improve slightly using the common item data (Table 24) for all of the rater scenarios by decreasing to 34 for one rater, 23 for two raters and 18 for four raters.

Table 21

D-study of Communications Dimension Using Ratings from Five Student Raters and SP

	Number of Cases			
Number of Raters	12	20	28	44
1	0.527	0.650	0.722	0.803
2	0.639	0.747	0.805	0.867
4	0.716	0.808	0.855	0.903

Table 22

Common Item D-study of Communications Dimension Using Ratings from Five Student Raters and SP

	Number of Cases			
Number of Raters	12	19	26	41
1	.542	.652	.720	.802
2	.651	.747	.802	.864
4	.723	.805	.850	.899

Table 23

D-study of Communications Dimension Using Ratings from Five Student Raters

	Number of Cases			
Number of Raters	12	19	25	38
1	0.563	0.671	0.729	0.803
2	0.661	0.755	0.802	0.860
4	0.723	0.805	0.845	0.859

Table 24

Common Item D-study of Communications Dimension Using Ratings from Student Raters

	Number of Cases			
Number of Raters	12	18	23	34
1	.593	.685	.736	.805
2	.682	.763	.805	.859
4	.738	.809	.844	.889

History Taking G-study

The results for the G-study on the history taking dimension can be seen in Tables 25 thru 28. Notice between Table 25 and 27 that the variance accounted for by rater drops from around 10% with the SP factored in to around 2% when the SP is removed. This is also reflected in the ICC values shown earlier. There was a very large effect for Medical

Student x Case in both models, indicating what is called "case specificity" in the OSCE literature. Looking at the common item data with the SP (Table 26), the variance accounted for goes from about 5% to less than 1% with the SP removed (Table 28).

Table 25

G-study of History Taking Dimensions Using Ratings from Five Student Raters and SP

Source	Variance Component	Relative Contribution
Medical Student	0.002959	7.90%
Rater	0.003641	9.72%
Case	0.006172	16.47%
Medical Student x Rater	-0.00034	0%
Medical Student x Case	0.015987	42.66%
Rater x Case	-0.00017	0%
Medical Student x Rater x Case	0.009219	24.60%

Table 26

Common Item G-study of History Taking Dimensions Using Ratings from Five Student Raters and SP

Source	Variance Component	Relative Contribution
Medical Student	0.00442	9.47%
Rater	0.002252	4.82%
Case	0.012266	26.27%
Medical Student x Rater	-0.00058	0%
Medical Student x Case	0.014853	31.81%
Rater x Case	0.00048	1.03%
Medical Student x Rater x Case	0.013	27.84%

Table 27

G-study of History Taking Dimensions Using Ratings from Five Student Raters

Source	Variance Component	Relative Contribution
Medical Student	0.003348	10.40%
Rater	0.000513	1.59%
Case	0.005998	18.63%
Medical Student x Rater	-0.00021	0%
Medical Student x Case	0.016081	49.96%
Rater x Case	-0.00024	-0.74%
Medical Student x Rater x Case	0.006702	20.82%

Table 28

Common Item G-study of History Taking Dimensions Using Ratings from Five Student Raters

Source	Variance Component	Relative Contribution
Medical Student	0.004269	9.54%
Rater	0.000433	0.97%
Case	0.014531	32.46%
Medical Student x Rater	-0.00058	0%
Medical Student x Case	0.016698	37.30%
Rater x Case	-0.00016	0%
Medical Student x Rater x Case	0.009571	21.38%

History Taking D-study

The results when using the history taking dimensions indicate a need for 35, 28, and 25 stations when using one, two, and four raters, respectively (Table 29). With common item data, the number of cases decreases by nine for one rater and eight for the other rater scenarios (Table 30). Factoring in the BARS training used on my student raters, the number of stations becomes 24 and 22 for using two and four raters, respectively, and drops to 28 stations when only using one rater. When using only the common items, the number of cases decreases by four and three cases for the four and two rater scenarios, and also drops by three to 25 cases for one rater. These results are summarized in Tables 29 thru 32.

Table 29

D-study of History Taking Dimension Using Ratings from Five Student Raters and SP

	Number of Cases			
Number of Raters	12	25	28	35
1	0.585	0.746	0.767	0.804
2	0.633	0.782	0.801	0.834
4	0.660	0.802	0.819	0.838

Table 30

Common Item D-study of History Taking Dimension Using Ratings from Five Student Raters and SP

	Number of Cases			
Number of Raters	12	17	20	26
1	.651	.726	.757	.802
2	.711	.777	.804	.842
4	.744	.805	.829	.863

Table 31

D-study of History Taking Dimension Using Ratings from Five Student Raters

	Number of Cases			
Number of Raters	12	22	24	28
1	0.638	0.764	0.779	0.804
2	0.674	0.791	0.805	0.828
4	0.693	0.806	0.819	0.841

Table 32

Common Item D-study of History Taking Dimension Using Ratings from Five Student Raters

	Number of Cases			
Number of Raters	12	18	21	25
1	.662	.746	.774	.803
2	.705	.782	.807	.833
4	.729	.801	.824	.849

Physical Examination G-study

The results of this G-study seen in Tables 33 thru 36 were not anticipated. All estimated variance components for the medical students were negative, indicating that the ratees contributed 0% of the overall variance. It appears the individual case drove the results of both models and this will be discussed more in depth in the discussion section.

Table 33

G-study of Physical Exam Dimensions Using Ratings from Five Student Raters and SP

Source	Variance Component	Relative Contribution
Medical Student	-0.00042	0%
Rater	0.006801	8.47%
Case	0.018329	22.84%
Medical Student x Rater	-7.6E-06	0%
Medical Student x Case	0.026807	33.40%
Rater x Case	0.008835	11.01%
Medical Student x Rater x Case	0.019917	24.82%

Table 34

Common Item G-study of Physical Exam Dimensions Using Ratings from Five Student Raters and SP

Source	Variance Component	Relative Contribution
Medical Student	-0.00143	0%
Rater	0.01132	7.98%
Case	0.009351	6.59%
Medical Student x Rater	-0.00061	0%
Medical Student x Case	0.064785	45.68%
Rater x Case	0.010616	7.49%
Medical Student x Rater x Case	0.047798	33.70%

Table 35

G-study of Physical Exam Dimensions Using Ratings from Five Student Raters

Source	Variance Component	Relative Contribution
Medical Student	-0.00099	0%
Rater	-0.00048	0%
Case	0.031018	38.60%
Medical Student x Rater	-0.00029	0%
Medical Student x Case	0.033864	42.14%
Rater x Case	0.001624	2.02%
Medical Student x Rater x Case	0.01561	19.43%

Table 36

Common Item G-study of Physical Exam Dimensions Using Ratings from Five Student Raters

Source	Variance Component	Relative Contribution
Medical Student	-0.00231	0%
Rater	-0.00046	0%
Case	0.019999	13.98%
Medical Student x Rater	-0.00038	0%
Medical Student x Case	0.087086	60.89%
Rater x Case	0.00159	1.11%
Medical Student x Rater x Case	0.037487	26.21%

Physical Examination D-study

Due to each medical student not contributing overall to their score on the clinical portion with 0% of the variance accounted for, a D-study was not meaningful because with zero variance due to medical student, increasing the number of scenarios will never result in a reliability of .80. However, in light of this development, the numbers from the MTMM below were used to extrapolate how many stations would be needed using one judge. This was done by calculating Cronbach's alpha for the PX portion of the test and then using the Spearman-Brown prophecy to determine the needed length to reach .80. The result of these calculations indicates that 19 stations are needed.

Validity of Ratings

MTMM. Computations for the first MTMM came from averaging the SP ratings and those of two raters (see the Appendix for the full design). The rating means were computed for each dimension on each case and a correlation matrix was computed. The resulting matrix is shown in Table 37 for all items and Table 38 for the common items. For Table 37, the average entry on the validity diagonal is .22, the average of the heterotrait-heteromethod entries is .12, and the average of the monomethod entries is .24. Thus, on average, the traits (communication, history, and physical exam) tend to correlate across cases more highly with the same traits than with different traits. The correlations for traits between cases are no higher than the correlations within the cases, however.

For the entries in Table 38, the average entry on the validity diagonal is .21, the average of the heterotrait-heteromethod entries is .09, and the average of the monomethod entries is .21. Thus, on average, the traits (communication, history, and physical exam) tend to correlate across cases more highly with the same traits than with different traits. The correlations between the cases are equal to the correlations within cases for the common items.

Table 37

Multi-trait Multi-method Matrix of SP and Raters

	RBCM	*RBHX*	*RBPX*	*SBCM*	*SBHX*	*SBPX*	*JLCM*	*JLHX*	*JLPX*	*JSCM*	*JSHX*	*JSPX*
RBCM	**0.25**											
RBHX	0.23	**0.78**										
RBPX	0.37	0.14	**0.82**									
SBCM	0.17	0.13	0.14	**0.29**								
SBHX	0.12	0.31	0.08	0.18	**0.73**							
SBPX	0.09	0.02	0.32	0.24	0.23	**0.83**						
JLCM	0.09	0.08	0.05	0.46	0.27	0.07	**0.34**					
JLHX	0.08	0.04	0.10	0.06	0.30	0.16	0.28	**0.58**				
JLPX	0.02	0.18	0.04	0.17	0.12	0.06	0.24	0.15	**0.31**			
JSCM	0.38	-0.02	0.16	0.25	0.08	0.01	0.22	0.04	-0.04	**0.30**		
JSHX	0.23	0.17	0.35	0.26	0.10	0.16	0.05	0.30	0.06	0.30	**0.82**	
JSPX	0.26	0.01	0.32	0.10	-0.01	0.26	0.12	0.26	0.14	0.26	0.32	**0.64**

Note: ICC of raters and SP are in bold on the diagonal. Validities are underlined

Table 38

Common Item Multi-trait Multi-method Matrix of SP and Raters

	RBCM	*RBHX*	*RBPX*	*SBCM*	*SBHX*	*SBPX*	*JLCM*	*JLHX*	*JLPX*	*JSCM*	*JSHX*	*JSPX*
RBCM	**0.25**											
RBHX	0.26	**0.78**										
RBPX	0.37	0.12	**0.81**									
SBCM	0.20	0.16	0.14	**0.29**								
SBHX	0.16	0.33	0.08	0.18	**0.75**							
SBPX	0.06	0.03	0.32	0.27	0.24	**0.82**						
JLCM	0.15	0.10	0.06	0.39	0.27	0.05	**0.24**					
JLHX	0.10	0.05	0.09	0.11	0.25	0.20	0.28	**0.70**				
JLPX	0.02	0.07	0.03	0.02	0.06	0.13	0.15	0.14	**0.40**			
JSCM	0.36	0.02	0.15	0.25	0.12	0.03	0.24	0.11	-0.07	**0.26**		
JSHX	0.20	0.16	0.33	0.25	0.04	0.07	0.06	0.29	-0.05	0.25	**0.74**	
JSPX	0.18	-0.07	0.19	0.03	-0.05	0.25	-0.07	0.08	0.13	0.15	0.11	**0.65**

Note: ICC of raters and SP are in bold on the diagonal. Validities are underlined

The second MTMM split the psychology raters and the SP and reanalyzed the data since the means of the ratings were different according to whether the rater was trained for the study or hired as an SP. These can be seen in Table 39 for all items and 40 for the common items. In this MTMM, the psychology raters' correlations appear below the diagonal and the SP correlations appear above the diagonal. For the psychology raters, the average entry on the validity diagonal is .19, the average of the heterotrait-heteromethod entries is .10, and the average of the monomethod entries is .19. Thus, on average, the traits (communication, history, and physical exam) tend to correlate across cases more highly with the same traits than with different traits. The correlations between the cases are equal to the correlations within cases for the common items. For the SPs, the average entry on the validity diagonal is .16, the average of the heterotrait correlations is .10, and the average of the monomethod correlations is .28. Thus dimensions within cases are slightly more correlated than dimensions across cases for the SPs. For the common items (Table 40), and for the psychology raters, the average entry on the validity diagonal is .19, the average of the heterotrait-heteromethod entries is .08, and the average of the monomethod entries is .16. Thus, on average, the traits (communication, history, and physical exam) tend to correlate across cases more highly with the same traits than with different traits. The correlations between the cases are slightly higher than the correlations within cases. For the SPs, the average entry on the validity diagonal is .05, the average of the heterotrait correlations is .10, and the average of the monomethod correlations is .22.

Table 39

Multi-trait Multi-method Matrix of SP Above Diagonal and Raters Below Diagonal

	RBCM	*RBHX*	*RBPX*	*SBCM*	*SBHX*	*SBPX*	*JLCM*	*JLHX*	*JLPX*	*JSCM*	*JSHX*	*JSPX*
RBCM	**0.52**	0.29	0.32	0.08	0.09	0.09	0.00	0.00	0.08	0.25	0.17	0.21
RBHX	0.14	**0.88**	0.24	0.14	0.18	-0.08	0.06	0.08	-0.01	0.07	0.19	0.09
RBPX	0.31	0.11	**0.88**	0.08	0.10	0.24	-0.12	-0.04	0.11	-0.01	0.32	0.34
SBCM	0.06	0.12	0.11	**0.52**	0.05	0.26	0.12	-0.05	0.27	0.09	0.33	0.14
SBHX	0.09	0.31	0.04	0.17	**0.93**	0.32	0.28	0.22	0.13	0.03	0.10	0.02
SBPX	0.00	0.04	0.28	0.18	0.16	**0.91**	0.05	0.12	0.24	-0.03	0.18	0.19
JLCM	0.03	0.07	0.17	0.43	0.18	0.07	**0.61**	0.34	0.24	0.11	0.07	0.12
JLHX	0.10	0.01	0.16	-0.01	0.29	0.17	0.21	**0.85**	0.33	0.00	0.14	0.13
JLPX	-0.02	0.19	0.00	0.11	0.07	0.06	0.13	0.07	**0.80**	-0.06	0.14	0.16
JSCM	0.34	-0.04	0.25	0.28	0.08	0.06	0.20	0.09	0.02	**0.58**	0.22	0.28
JSHX	0.19	0.11	0.32	0.17	0.10	0.12	-0.01	0.32	-0.02	0.31	**0.82**	0.39
JSPX	0.21	-0.07	0.28	0.08	-0.04	0.26	0.03	0.22	0.14	0.29	0.24	**0.73**

Note: ICC of my raters is in bold on the diagonal. Validities are underlined.

Table 40

Common Item Multi-trait Multi-method Matrix of SP Above Diagonal and Raters Below Diagonal

	RBCM	*RBHX*	*RBPX*	*SBCM*	*SBHX*	*SBPX*	*JLCM*	*JLHX*	*JLPX*	*JSCM*	*JSHX*	*JSPX*
RBCM	**0.48**	0.30	0.31	0.04	0.09	0.07	-0.01	-0.06	-0.05	0.26	0.18	0.07
RBHX	0.18	**0.88**	0.24	0.13	0.18	-0.08	0.08	0.12	-0.08	0.07	0.23	-0.04
RBPX	0.31	0.09	**0.85**	0.06	0.10	0.24	-0.11	-0.13	0.16	-0.01	0.29	0.18
SBCM	0.16	0.14	0.12	**0.51**	-0.01	0.25	0.09	0.01	0.13	0.07	0.35	0.00
SBHX	0.13	0.33	0.04	0.17	**0.92**	0.32	0.25	0.15	-0.01	0.03	-0.01	-0.09
SBPX	-0.03	0.07	0.29	0.22	0.18	**0.91**	0.03	0.08	0.15	-0.03	0.04	0.12
JLCM	0.11	0.08	0.15	0.39	0.16	0.04	**0.47**	0.33	0.23	0.11	0.10	0.03
JLHX	0.15	-0.01	0.19	0.07	0.24	0.23	0.20	**0.79**	0.26	0.07	0.12	0.00
JLPX	0.02	0.08	-0.02	0.00	0.02	0.15	0.09	0.06	**0.80**	-0.01	0.15	0.04
JSCM	0.31	0.00	0.22	0.29	0.11	0.06	0.22	0.11	0.07	**0.57**	0.25	0.09
JSHX	0.12	0.08	0.27	0.10	0.04	0.04	-0.01	0.29	-0.12	0.22	**0.77**	0.12
JSPX	0.15	-0.15	0.18	0.01	-0.07	0.25	-0.06	0.03	0.20	0.22	0.05	**0.84**

Note: ICC of my raters is in bold on the diagonal. Validities are underlined.

Overall, the MTMM analyses show some evidence of convergent validity. However, the correlations of the same traits (communication, history taking, and physical examination) over cases are not large, either for the psychology raters or the SPs. There is evidence for discriminant validity, however, as the entries in the validity diagonal are consistently larger than the other entries in the heteromethod blocks. The entries in the validity diagonal tend to be about equal to those in the monomethod triangles. Thus, neither trait effects (dimensions) nor method effects (cases or exercises) appear dominant in these MTMM matrices. It is also worth noting that there was no clear pattern to the size of the correlations and that the alternate forms of exercises do not show consistently larger correlations than different forms. That is, the agreement between ratings for RB and SB and between JL and JS are not much different than agreement between RB and JL or SB and JS.

Factor Analysis. Exploratory factor analysis was used to achieve a second perspective on the validity of the ratings data. Extraction of the factors was done using the maximum likelihood method and the data were rotated orthogonally due to the *a priori* assumption that the three dimensions (CM, HX, and PX) should not be correlated.

Table 41

EFA Restricted to Three Factors

	Factor		
	1	2	3
rbpx	.627	.072	.076
jshx	.528	.099	.228
jspx	.464	.123	.194
sbpx	.423	.168	-.059
rbcm	.411	.077	.324
rbhx	.239	.161	-.061
jlcm	-.080	.839	.173
sbcm	.162	.501	.196
sbhx	.174	.357	.033
jlhx	.268	.348	-.021
jlpx	.093	.323	-.072
jscm	.135	.066	.988

Extraction Method: Maximum Likelihood.
Rotation Method: Varimax with Kaiser Normalization.

The scree plot indicated a three factors solution. The first factor contains scales for all three content areas (physical exam and history taking, and communication). The second factor appears highly related to communication but only for John Long. The third factor is closely associated with communication for John Sexton. Thus, the factors do not correspond neatly to either dimensions or methods, and the best interpretation is unclear. In an exam such as this one, it would be desirable that the dimensions of communication (CM), history taking (HX), and physical examination (PX) would make up the three factors and that the cases of RB, SB, JL, and JS would not have much of an effect. However, the desired and actual results are not in close correspondence. Recall that the

RB and JS examination took place at the same time and the SB and JL exam took place together. It is not clean by any measure, but I would speculate that the individual student's differing skill levels are responsible for the EFA results. The reason the results are not clearer may be due to different students progressing at different levels, however, overall, the whole sample of students would be increasing their skill level over time.

Chapter Four

Discussion

The psychometric quality of assessments of medical student clinical competence is important for both theoretical and practical reasons. Reliability of evaluations is essential for feedback to individuals and for fairness and utility of decisions regarding individual proficiency. Understanding the sources of variance in performance measures is helpful for designing performance examinations and for designing the training needed to ensure that medical practitioners are properly skilled.

The generalizability analysis provided information about the relative importance of raters, cases, and roles within case as sources of variance in assessments of student competence. The analyses also allowed us to describe the reliability of (part of) the current CPX and to estimate the reliability of hypothetical future exams. Such information can help ensure quality exam plans in the future. It can also help us to understand why examination performance is less than perfectly reliable and the anticipated realistic upper limits for assessment reliability.

The MTMM analyses allowed us to examine the convergent and discriminant validity of scores on an existing CPX, which were analyzed both using the Campbell-Fiske criteria and factor analysis. The validity of ratings is especially important for the evaluation of technical versus interpersonal skill in clinical competence. Patients are ordinarily not adept at assessing technical competence of doctors, but they can and do assess the quality of interpersonal relations in doctor-patient encounters. This study allowed us to assess both technical and interpersonal skills over several different contexts and to determine the degree to which they are discriminable from one another and the degree to which they converge over situations. We were able to assess the effect of similarity of situation on the convergent and discriminant validity by including multiple cases, some of which could be considered alternate forms.

Validity of Measurement

Campbell and Fiske (1959) suggested several criteria for evaluating the validity of measures using the MTMM matrix. For convergent validity, one should find that the entries in the validity diagonal are large enough to be statistically significant and practically meaningful. Additionally, one should find that the entries in the validity diagonal are larger than the corresponding entries in the heteromethod blocks. For the factor analysis, one would hope to see factors that correspond to the traits (communication, history taking, and physical exam). Although there was some evidence for convergent validity, the evidence was not as strong as one would like. The entries in the validity diagonals were often large enough to be statistically significant, and were, on average, larger than the relevant comparisons in the heteromethod blocks. However, there were also some very small convergent validities (especially comparing scores from the Rachel Brown and John Long scenarios), and none of the entries was large (the largest convergent correlation was .46, for the average overall ratings for communication between Samantha Browning and John Long; see Table 37).

An additional point to note is that the exercises were paired into two sets of two cases. If there are certain characteristics of physicians related to the way in which they deal with patients in terms of gathering information and diagnostic reasoning, we would expect to see largest convergent correlations within cases rather than between cases. If the physicians have obvious biases in dealing with patients on the basis of sex, we would expect to see larger convergent correlations within case as well (because cases will be either male or female). However, there was no obvious pattern in which the correlations were larger within cases than between cases. Such a finding is important because at least for the cases studied here, the details of the case do not seem to be driving the evaluations.

Although this was not hypothesized, it appears the time the exams were taken had an effect on convergent validity. Overall the cases that took place in the fall, RB and JS, had better convergent validity and the spring cases, SB and JL, had better convergent validity. Greater correlations for performance measures taken closer in time is not new, and generally indicates that individuals differ in their growth rates for competence. That is, people move up and down the performance distribution in relation to the average person relatively slowly, so that relative orderings of people on performance tend to be larger as they are taken closer in time.

I was surprised that RB and SB did not have better convergent validity on the history taking (*r* = *.31*) and physical examination (*r* = *.32*) dimensions, since they were the exact same items. This may have been due to a lack of variability, which will attenuate correlations. JS and JL faired almost as well *(r = .30)* on history taking as RB and SB, but not nearly as well on HX (*r* = *.14*). The lack of convergent validity for the JL on the dimension of physical examination may have been the result of a floor effect. The average score on this section of the exam, according to the psychology raters was a 23.7%. Since there were only four items in the clinical examination section of the JL case, this indicates that the average score was less than one item correct. Reasons for this low score are most likely due to the specificity of the items (e.g., Did the student listen to your heart under the gown?). Most of the students performed the action that would usually result in an endorsement of the item (listening to the heart), but they did not do the action exactly as specified on the rating sheet (e.g., under the gown), therefore, they did not get credit for it. Also, the SP was much more likely to give the student credit for checking the heart under the gown, when in fact the medical student had not done so. Possible reasons for this will be discussed later.

Regarding discriminant validity, we would like to see correlations among traits that are rather small, especially within methods (correlations found in the monomethod blocks). We would not like to see factors that correspond to methods. For this type of matrix (where simulation exercises are considered methods), the correlations were rather low, ranging from the teens through the .30s.

When compared to the MTMM matrix computed by Bycio, Alvares, and Hahn (1987) on Assessment Center ratings, the discriminant validity of these dimensions looks very good. In the Bycio et al. (1987) study, the heterotrait-heteromethod correlations were frequently larger than .60, while for this simulation, they stayed under .40. Another study that looked at convergent and discriminant validity of an assessment center, Arthur, Woehr, and Maldegen (2000), showed better convergent validity than this study, but again discriminant validity was not as good as what was found in the current study. Structuring an assessment center in the business world may benefit from more exercises for shorter durations, mirroring the OSCE structure.

There is some evidence for case specificity or exercise effects in the results of the present study. But, when we look at the two studies of assessment center validity previously mentioned, the off diagonal correlations in this study are much lower than those found in the assessment center environment of Bycio et al. (1987) and Arther et al. (2000). This may be due to the use of checklists instead of rating scales.

MTMM matrices were computed using both all the items and also just the common items across scenarios. I expected the MTMMs based on the common items to look better than the MTMMs for all the items because of their identical content. However, the MTMM based on the common items did not look any better than the matrix based on all the items. This result may be partly an artifact of reliability, meaning that the shorter scales' lower reliability offset the improvement in convergence due to identical content.

Reliability of Measurement

Reliability of measurement is fundamental to theoretical development and to practical application. The generalizability analysis provided information about the magnitude of several sources of variability in measures overall and for each of the three dimensions of interest in this study (communication, history taking, physical exam). For the ICC values, keep in mind that since this was done assuming random instead of fixed raters, there are some slight differences between the ICC values of the standard exam and the common item analysis where they should, theoretically, be the same. This occurs for the cases of RB and SB on the dimension of HX and PX.

Ideally, the vast majority of variance in measures can be attributed to individual differences in the ratees, in this case, to the differences in skill of medical students in communication, history taking and giving of physical examinations. Variance in measurement due to cases, raters, and other factors represents nuisance when the objective of the measurement is to evaluate trainee skill. Reliability of measurement was examined by estimating variance components, estimating the reliability of measures as they are currently used, and projecting what reliability would be using hypothetical raters and examinations, both for overall assessments and for separate dimensions.

Reliability of Scales in Current Use. Table 8 shows the estimates of reliability for each exercise and dimension for a single, random judge. The estimates including the SPs are uniformly low, but this is doubtless due in large part to the differences in means between the psychology raters and the SPs. The data from just the psychology raters can be used as an alternative estimate of the reliability of the current scoring system. Based on these ratings, the history taking measures show adequate or better reliability for each exercise. The physical exam scores show good reliability for some exercises, but not others. This is likely due to some exercises having a small number of scored items or ceiling/floor effects for the particular case. The communication scales showed poorer reliability across exercises.

Reliability of Future Examinations. Generalizability theory was used to estimate the reliability of hypothetical examinations similar to the current exam, but composed of various numbers of raters (judges, SPs) and cases or exercises. One hopes to reach a reliability of .80 for the overall exam and on individual dimensions with between 12 cases, as is the current format, or as many as 16, which would still be a practical number. Instead, it took many more cases than expected, with some dimensions requiring over 40 to reach the threshold with one rater.

Overall. Here, when psychology raters were analyzed with the SP, the medical student accounts for around 6% of the true score variance and the rater accounts for about 35%. When the SP is removed, these numbers change to 11% of the true score variance by the student and drop to only 9% accounted for by the raters. For both studies, a large percentage, 20% with the SP and 32% without the SP, was accounted for by the interaction between student and case. The latter result indicates that students do better on some cases than others. Such a result is consistent with the literature on the OSCE. The effect is called "case specificity," which refers to the tendency for physicians to receive high marks on some cases, but relatively low marks on others. This effect is not simply due to some cases being more difficult than others; as such an effect would appear in the variance component for cases (the main effect for cases rather than the interaction between cases and students).

Communication. The Communication dimension of the exam was by far the most subjective dimension. Unlike the other scales, which were measured with checklists, the communication dimension was measured with Likert type scales (summated rating scales). Behaviorally anchored rating scales were developed for Communication and coupled with training on using the scales in order to maximize agreement among the psychology raters. The results of the G-study on communication with my five raters and the SP showed only 4% of the variance accounted for by the medical student and a very large percent (44), accounted for by the raters. Interestingly, only 11% was accounted for by the interaction between student and case. This is quite different from the results Guiton et al. (2004) had with their G-study on communication. In their model, 50.16% of the variance was attributable to students by case. Keep in mind that there are several differences here including number of cases, four in mine vs. seven in theirs and they had a sample of over 300 students, but these are still large differences.

When the SP is removed from the ratings and just my raters results are used, the BARS training makes an impact. The variance accounted for by rater plummets from 44% down to 13%. Students now account for 7%, which is slightly better than the 4% with the SP included. Medical student by case interaction rises to just over 20% indicating more variance by student from case to case.

History taking. These results look better all around. With the SP and my raters, students account for about 8% of the true score variance, while raters only accounted for fewer than 10%, indicative of much more consistency. Recall that history taking was measured using a “yes or no” checklist. The student by case interaction accounted for about 43% of the variance, indicating students performed differently from case to case. When the SPs ratings were removed, the medical students’ variance increased slightly to just over 10%. Rater variance was down to less than 2%, which is excellent. However, student by case variance increased to 50%, indicating students performed much differently on their history taking skills from case to case.

Physical examination. The results when using the physical examination dimensions were problematic for the D-study due to the variance component estimates of zero. The unusual result may be due to range restriction in the case of JL where there were only 4 items. The ratings for JL appear subject to a floor effect. Other results consistent with the floor effect include nearly 39% of the variance being attributable to case and 42% of the variance being attributable to the student by case interaction. When the SP rater is removed, the results are similar. However, the alternate analysis of reliability using the Spearman-Brown prophecy based on correlations from the MTMM indicated that 19 cases would be necessary to achieve an alpha reliability of .80.

Using the common item data did not change the results of the G and D-studies as much as expected. I was hoping that by reanalyzing using the common data, we would be able to pull some variance accounted for out of just the medical students for the physical examination dimension, but that did not happen. Again, using only the common items between the cases of JL and JS shorted the scale to only two items which exacerbated the floor effect that was already taking place.

Psychology raters vs. SPs

The SPs consistently awarded higher scores to the medical students than did the psychology raters. This occurred both for the more subjective communication items as well as the less subjective checklist items. Based on conversations with administrators in the testing facility, it appears that some SPs develop bonds with the medical students, sincerely wish them well, and are apparently rather generous in their marks. In conversations with SPs regarding communication, some SPs assume that the medical student will be given the highest marks unless they do something out of the ordinary, such as making rude comments or simply failing to communicate at all. There may be other reasons for the differences, but the differences are large enough to warrant further attention.

Dimensions and formats

Although checklists help provide evaluations that are relatively objective, they may also be deficient. For example, if a patient has a missing leg, and a physician fails to ask about it, one would doubt that physician's history taking skill. However, if the checklist has no item regarding the leg, this oversight will not be counted. Increasing the length of checklists has also been criticized in the literature as adding irrelevant variance. The subjective assessment allows the judge or rater to include all the essential behaviors in the evaluation, but the choice of behaviors and the determination of what is essential is left up to the judge, thus allowing for quite a bit of difference due to judge idiosyncrasies. In this study, therefore, it was possible that the judges might show less agreement within cases (e.g., lower reliability for RB, for example), but higher reliability between cases (e.g., higher reliability between RB and SB, for example), in comparison to the dimensions assessed by checklist. In other words, we might expect better within case agreement but worse between case agreement for checklists as compared to rating scales. However, such was not the case in the current study. Reliability for communication was lower within cases, but essentially the same between cases (on average) as history and physical exam scores.

An analysis of individual psychology rater data (see Appendix C) shows that correlations between cases do not appear higher when rated by individuals than when averaged across raters. It could be, for example, that some raters are sensitive to specific interpersonal cues that sway their judgments of communication competence. If physicians consistently display such cues and raters consistently use them in their analyses, we might see high correlations between cases for individual judges, but not for the average of the judges. Such was not generally the case, however, as the correlations between cases appeared similar in magnitude for individual judges and for the average judge. Rater 3 is a possible exception, showing relatively large positive correlations among all the communications ratings, but not the history taking and physical examination dimensions. Reasons for this are unknown as this rater was trained and monitored the same as the other four psychology raters.

Comparison to other medial simulation studies

Overall this study differs from many of the studies listed in the introduction section because our communication reliabilities were much lower than those found in other studies (Brailovsky & Grand'maison, 2000); Amiel et al., 1997). The g-studies showed that our dimensions were not as reliable as others had found for comparable dimensions like history taking (Lee, Wilkinson, Battles, & Hynan, 2003). Additional differences include our study having alternate forms of a case where the SP is played by an actor and also a mannequin. As far as raters go, this is the only study I know about that has used raters from outside of the medical department. Training these raters was fairly straight forward and not very time consuming. An advantage to using raters from outside the department is that they will most likely be very objective in their ratings since they have nothing to gain or lose deepening on how the medical students are scored. Involving raters from a different department may also work towards building a bridge between two (or more) disciplines that may spark the embers of future symbiotic research.

Limitations

Limitations of this study include a small sample size of 20 for the G-study/D-study calculations. This resulted from the tremendous time investment required to get data of this nature. To rate the nearly 300 videos required over 40 hours of rating per rater, not including filling out the rating sheets and putting the ratings into the database. Thus, although each medical student at each station was rated by an SP and two psychology raters, the two psychology raters were not the same two for each student and station. In other words, psychology student raters were spread systematically across stations and students. This limitation was partially offset by comparing estimates from the whole study (ignoring particular rater effects) to results from the subset of 20 students in which particular rater effects could be properly estimated.

A strength of the study is that actual medical students were studied during performances that were graded (i.e., the performance counted) so that the results of the study are likely to be generalizable to medical examinations using standardized patients. However, the sample was from a single medical school using a small number of professional standardized patients. It is unknown how well the current results may generalize to other comparable settings.

Implications for Medical Testing

There are several implications for the future of medical testing that can be garnered from this study. First, the accuracy of the scores cannot be taken for granted. Anecdotally, it has come to my attention that when one is employed as a SP, it tends to be a reoccurring job. Once a SP has been successfully trained, he or she tends to get utilized for several iterations of the different medical examinations for which their particular ailment is included. Because of this, the SPs may build bonds between themselves and the medical students which may lead to leniency on the scoring of exams. Also, there is a tremendous amount of pressure on students while going through medical school. Many SPs may feel uncomfortable grading the medical students too harshly, even if that grade is more accurate. I believe that my raters having zero ties to the medical college helped them to be very objective when rating the medical students and think that this may be the main reason for such large differences on all dimensions and across all cases between my raters' ratings and the SP's ratings. Another factor that may have contributed to the differences in scores is that my raters rated the videos as the examination happened. The SPs, due to the nature of the exam, had to wait until the examination was over to enter their ratings. It is difficult to remember precisely what happened during an examination that lasted around 10 minutes. This is especially true if you are seeing 20 medical students (playing doctors) in a row. If the SP could not remember for certain if a medical student completed an item, the SP may have chosen to give the student the benefit of the doubt.

Another area that we have touched on with this study is checklists vs. more general ratings. The evidence suggests that it is risky to use a single judge to assess communication. Communication is more reliably assessed when using multiple judges (at least two).

According to the current results, a single exam composed of cases similar to those studied here would need to be considerably longer than the current exam in order to achieve a reliability of greater than .80. It might be worthwhile to examine the scores of current students across all cases (all scores in the OSCE or CPX) to estimate the reliability of the overall exam. Should that estimate also be lower than desired, the feasibility of longer exams should be considered in light of the medical school's goals.

The medical school may want to look at training the SP with the BARS in order to increase the reliability of the communications dimension. Periodic checkups where a Subject Matter Expert (SME) reviews the interactions of SPs and medical students via videotape and then prescribes additional training if needed could also boost the reliability of the dimensions graded by checklist.

Future directions

Assessments of interpersonal communication (social skill, interpersonal competence) may not be measured well with only a few interpersonal interactions. If we want good assessments of communication skill, perhaps we should include interactions with a broad array of people, including different demographic groups (race, sex, age) as well as some communication difficulties (e.g., language barriers, deafness, high anxiety, etc.). Future research should investigate devising good stimulus materials (i.e., scenarios and associated SPs) for the assessment of communication.

Future research also appears warranted for the training of SPs, particularly for the more subjective aspects of physician performance. It may be necessary to evaluate or grade SPs on their own evaluative performance, perhaps by checking the SP scores against expert scores based on a video recording of the physician-patient encounter. Another possibility is to create some "standardized doctors", so that the SP would see a physician acting in a manner designed to produce a particular score. Such encounters might also serve as a periodic check on SP accuracy and/or continued training. Another apect to look at is if SPs rate differently through out the day. Fatigue may set in after seeing a number of medical students and this could affect ratings.

Future research is needed to better understand the reasons for the ubiquitous "case specificity" effect. Because the effect is found in diverse settings of evaluations of skilled performance (e.g., the exercise effect in the assessment center as well as the OSCE), and the effect is found both using the relatively objective checklist as well as the relatively subjective Likert scale, it appears to be something beyond a simple effect of measurement or evaluation. In other words, the effect is not just in the head of the judge; people really do respond better to some situations than to others, even though the situations are thought to be equivalent by the test developers. The reasons for such an effect are not obvious, other than to note the “fundamental attribution error,” which suggests that behavior is far more attributable to minor extraneous situational influences than we think.

References

Amiel, G. E., Tann, M., Krausz, M. M., Bitterman, A., & Cohen, R. (1997). Increasing examiner involvement in an objective structured clinical examination by integrating a structured oral examination. *American Journal of Surgery, 173(6),* 546-549.

Amiel, G. E., Ungar, L., Alperin, M., Baharier, Z., Cohen, R., & Reis, S. (2006). Ability of primary care physician's to break bad news: A performance based assessment of an educated intervention. *Patient Education and Counseling*, *60*, 10-15.

Arthur Jr., W., Woehr, D. J., & Maldegen, R. (2000). Convergent and Discriminant Validity of Assessment Center Dimensions: A Conceptual and Empirical Reexamination of the Assessment Center Construct-Related Validity Paradox. *Journal of Management*, *26*(4), 813-835.

Battles, J. B., Wilkinson, S. L., & Lee, S. J. (2004). Using standardised patients in an objective structured clinical examination as a patient safety tool. *Quality and Safety in Health Care*, *13*, 46-50.

Brailovsky, C. A., & Grand'maison, P. (2000). Using evidence to improve evaluation: A comprehensive psychometric assessment of a SP-based OSCE licensing examination. *Advances in Health Sciences Education*, *5*, 207-219.

Brannick, M. T., Michaels, C. E., & Baker, D. P. (1989). Construct validity of in-basket scores. *Journal of Applied Psychology*, *74*, 957-963.

Bycio, P., Alvares, K. M., & Hahn, J. (1987). Situational SPecificity in Assessment Center Ratings: A Confirmatory Factor Analysis. *Journal of Applied Psychology*, *72*(3), 463-474.

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

Carraccio, C., & Englander, R. (2000). The objective structured clinical examination. *Archives of Pediatrics and Adolescent Medicine*, 736-741.

Cohen, R., Rothman, A. I., Ross, J., & Poldre, P. (1991). Validating an objective structured clinical examination (OCSE) as a method for selecting foreign medical graduates for a pre-internship program. *Academic Medicine*, *66*, S67-S69.

Guiton, G., Hodgson, C. S., 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.

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

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

Hubbard, J. P., Levit, E. J., Schumacher, C. F., & Schnabel, T. G. (1965). An objective evaluation of clinical competence. *New England Journal of Medicine*, *272*, 1321-1328.

Hutchinson, L., Aitken, P., & Hayes, T. (2002). Are medical postgraduate certification processes valid? A systematic review of the published evidence. *Medical Education*, *36*, 73-91.

Larsen, T., & Jeppe-Jensen, D. (2008). The introduction and perception of an OSCE with an element of self- and peer-assessment. *European Journal of Dental Education*, *12*, 2-7.

Mann, K. V., Macdonald, A. C., & Norcini, J. J. (1990). Reliability of objective structured clinical examinations: four years of experience in a surgical clerkship. *Teaching and Learning in Medicine*, *2*, 219-224.

Matsell, D. G., Wolfish, N. M., & Hsu, E. (1991). Reliability and validity of the objective structured clinical examination in pediatrics. *Academic Medicine*, *25*, 293-299.

Mavis, B., Henry, R., Ogle, K., & Hoppe, R. (1996). The Emperor's new clothes; The OSCE revisited. *Academic Medicine*, *71*, 447-453.

Minion, D. J., Donnelly, M. B., Quick, R. C., Pulito, A., & Schwartz, R. (2002). Are multiple objective measures of student performance necessary? *The American Journal of Surgery*, *183*, 663-665.

Nayer, M. (1993). An overview of the objective structured clinical examination. *Physiotherapy Canada*, *45*, 171-178.

Neidig, R. D., & Neidig, P. J. (1984). Multiple assessment centers exercises and job relatedness. *Journal of Applied Psychology*, *69*, 182-184.

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

Newble, D., Dauphinee, D., Dawson-Saunders, B., MacDonald, M., Mulholland, H., Page, G., Swanson, D., Thomson, A., & Van Der Vleuten, C. (1994). Guidelines for the development of effective and efficient procedures for the assessment of clinical competence. In D. Newble, B. Jolly, & R. Wakeford (Eds.), *The Certification and Recertification of Doctors: Issues in the Assessment of Clinical Competence.* (pp. 69-91). Cambridge: Cambridge University Press.

Newble, D., Hoare, J., & Sheldrake, P. F. (1980). The selection and training of examiners for clinical examinations. *Medical Education*, *14*, 345-349.

Newble, D. L., & Swanson, D. B. (1983). Psychometric characteristics of the objective structured clinical examination. *Medical Education*, *22*, 325-334.

Norman, G. R. (1985). Objective measurement of clinical performance. *Medical Education, 19(1),* 43-47.

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

Olson, L. G. (1999). The ability of a long-case assessment in one discipline to predict students' performance on long case assessments in other disciplines. *Academic Medicine*, *74*, 835-839.

Park, R. S., Chibnall, J. T., Blaskiewicz, R. J., Furman, G. E., Powell, J. K., & Mohr, C. J. (2004). Construct validity of an objective structured clinical examination (OSCE) in psychiatry: Associations with the clinical skills examination and other indicators. *Academic Psychiatry*, *28*, 122-128.

Petrusa, E. R., Guckian, J. C., & Perkowski, L. C. (1984). A multiple station objective clinical evaluation. *Proceedings of the Twenty-third Annual Conference on Research in Medical Education, 23,* 211-216.

Petrusa, E. R. (2002). Clinical Performance Assessment. In G. R. Norman, C. P. Van Der Vleuten, & D. I. Newble (Eds.), *International Handbook of Research in Medical Education* (pp. 673-709). Dordrecht, Great Britain: Kluwer Academic Publishers.

Quest, T. E., Ander, D. S., & Ratcliff, J. J. (2006). The validity and reliability of the affective competency score to evaluate death disclosure using standardized patients. *Journal of Palliative Medicine*, *9*, 361-370.

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

Rutala, P. J., Witzke, D. B., Leko, E. O., & Fulginiti, J. V. (1990). The influence of student and standardized patient genders on scoring in an objective structured clinical examination. *Academic Medicine*, *66*, S28-S30.

Sackett, P. R., & Dreher, G. F. (1982). Constructs and assessment centers dimensions: Some troubling empirical findings. *Journal of Applied Psychology*, *67*, 401-410.

Schneider, J. R., & Schmitt, N. (1992). An exercise design approach to understanding assessment center dimension and exercise constructs. *Journal of Applied Psychology*, *77*, 32-41.

Schuwirth, L., & Van Der Vleuten, C. (2003). The use of clinical simulations in assessment. *Medical Education (Suppl.)*, *37*, 65-71.

Schwartz, R. W., Witzke, D. B., Donnelly, M. B., Stratton, T. S., Blue, A. V., & Sloan, D. A. (1998). Assessing residents' clinical performance: Cumulative results of a four-year study with the objective structured clinical examination. *Surgery*, *124*, 307-312.

Shatzer, J. H., Darosa, D., Colliver, J. A., & Barkmeier, L. (1993). Station-length requirements for reliable performance-based examination scores. *Academic Medicine*, *68*, 224-229.

Shatzer, J. H., Wardrop, J. L., Williams, R. G., & Hatch, T. F. (1994). The generalizability of performance on different-station-length standardized patient cases. *Teaching and Learning in Medicine*, *6*, 54-58.

Sibbald, D., & Regehr, G. (2003). Impact on the psychometric properties of a pharmacy OSCE: Using 1st-year students as standardized patients. *Teaching and Learning in Medicine*, *15*, 180-185.

Sloan, D., Donnely, M., Schwartz, R., & Strodel, W. (1995). The objective structured clinical examination: the new gold standard for evaluating post graduate clinical performance. *Annals of Surgery*, *22*, 735-742.

Smith, L. J., Price, D. A., & Houston, I. B. (1984). Objective structured clinical examination compared with other forms of student assessment. *Archives of Disease in Childhood*, *59*, 1173-1176.

Stillman, P. L., Regan, M. B., Swanson, D. B., Case, S., McCahan, J., Feinblatt, J., Smith, S. R., Williams, J., & Nelson, D. V. (1990). An assessment of the clinical skills of fourth-year students at four New England medical schools. *Academic Medicine*, *65*, 320-326.

Swanson, D. B. (1987). A measurement framework for performance based tests. In I. R. Hart, & R. M. Harden (Eds.), *Further Developments in Assessing Clinical Competence* (pp. 13-45). Montreal: Can-Heal.

Swanson, D. B., Clauser, B. E., & Case, S. M. (1999). Clinical skills assessment with standardized patients in high-stakes tests: a framework for thinking about score precision, equating, and security. *Advances in Health Sciences Education*, *4*, 67-106.

Tombleson, P., Fox, R. A., & Dacre, J. A. (2000). Defining the content for the objective structured clinical examination component of the Professional and Linguistic Assessments Board examination: development of a blueprint. *Medical Education*, *34*, 566-572.

Vu, N. V., Barrows, H. S., March, M. L., Verhulst, S. J., Colliver, J. A., & 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*, 43-50.

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

Wass, V., Jones, R., & Van Der Vleuten, C. (2001). Standardized or real patients to test clinical competence? The long case revisited. *Medical Education*, *35*, 321-325.

Waterson, T., Cater, J. I., & Mitchell, R. G. (1980). An objective undergraduate clinical exam in child health. *Archives of Disease in Childhood*, *55*, 917-922.

Watson, A. R., & Houston, I. B. (1982). Evaluation of an objective structured clinical examination. *Archives of Disease in Childhood*, *57*, 390-398.

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

Wilkinson, T. J., & Fontaine, S. (2002). Patients' global ratings of student competence. Unreliable contamination or gold standard? *Medical Education*, *36*, 1117-1121.

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

Wilson, G. M., Lever, R., Harden, R. M., Robertson, J. I., & Macritchie, J. (1969). Examination of clinical examiners. *Medical Education*, *1*, 37-40.

Appendices

Appendix A

Rating Diagram

Student	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Number
1a	X	X				1
1b	X		X			
1c	X			X		
1d	X				X	
2a		X	X			2
2b		X		X		
2c		X			X	
2d	X	X				
4a			X	X		3
4b			X		X	
4c	X		X			
4d		X	X			
5a				X	X	4
5b	X			X		
5c		X		X		
5d			X	X		
6a	X	X	X	X	X	5
6b	X	X	X	X	X	
6c	X	X	X	X	X	
6d	X	X	X	X	X	
7a				X	X	6
7b			X		X	
7c		X			X	
7d	X				X	
8a			X	X		7
8b		X		X		
8c	X			X		
8d				X	X	
9a		X	X			8
9b	X		X			
9c			X		X	
9d			X	X		
11a	X	X				9
11b		X			X	
11c		X		X		
11d		X	X			
12a	X	X	X	X	X	10
12b	X	X	X	X	X	
12c	X	X	X	X	X	
12d	X	X	X	X	X	
14a	X				X	11
14b	X			X		

Appendix A (continued)

Student	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Number
14c	X		X			
14d	X	X				
16a	X				X	12
16b	X			X		
16c	X		X			
16d	X	X				
17a	X				X	13
17b	X			X		
17c	X		X			
17d	X	X				
18a	X	X				14
18b		X			X	
18c		X		X		
18d		X	X			
19a		X	X			15
19b	X		X			
19c			X		X	
19d			X	X		
22a			X	X		16
22b		X		X		
22c	X			X		
22d				X	X	
24a				X	X	17
24b			X		X	
24c		X			X	
24d	X				X	
26a	X				X	18
26b		X			X	
26c			X		X	
26d				X	X	
27a				X	X	19
27b	X			X		
27c		X		X		
27d			X	X		
28a			X	X		20
28b			X		X	
28c	X		X			
28d		X	X			
31a	X	X	X	X	X	21
31b	X	X	X	X	X	
31c	X	X	X	X	X	
31d	X	X	X	X	X	
33a		X	X			22
33b		X		X		

Appendix A (continued)

Student	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Number
33c		X			X	
33d	X	X				
34a	X	X	X	X	X	23
34b	X	X	X	X	X	
34c	X	X	X	X	X	
34d	X	X	X	X	X	
35a	X	X				24
35b	X		X			
35c	X			X		
35d	X				X	
36a		X	X			25
36b		X		X		
36c		X			X	
36d	X	X				
38a			X	X		26
38b			X		X	
38c	X		X			
38d		X	X			
40a	X	X	X	X	X	27
40b	X	X	X	X	X	
40c	X	X	X	X	X	
40d	X	X	X	X	X	
41a	X				X	28
41b		X			X	
41c			X		X	
41d				X	X	
43a				X	X	29
43b			X		X	
43c		X			X	
43d	X				X	
46a			X	X		30
46b		X		X		
46c	X			X		
46d				X	X	
47a	X	X	X	X	X	31
47b	X	X	X	X	X	
47c	X	X	X	X	X	
47d	X	X	X	X	X	
48a		X	X			32
48b	X		X			
48c			X		X	
48d			X	X		
49a	X	X				33
49b		X			X	

Appendix A (continued)

Student	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Number
49c		X		X		
49d		X	X			
50a	X				X	34
50b	X			X		
50c	X		X			
50d	X	X				
51a	X				X	35
51b	X			X		
51c	X		X			
51d	X	X				
54a	X	X				36
54b		X			X	
54c		X		X		
54d		X	X			
55a		X	X			37
55b	X		X			
55c			X		X	
55d			X	X		
56a			X	X		38
56b		X		X		
56c	X			X		
56d				X	X	
57a				X	X	39
57b			X		X	
57c		X			X	
57d	X				X	
58a	X				X	40
58b		X			X	
58c			X		X	
58d				X	X	
59a	X	X	X	X	X	41
59b	X	X	X	X	X	
59c	X	X	X	X	X	
59d	X	X	X	X	X	
60a				X	X	42
60b	X			X		
60c		X		X		
60d			X	X		
61a			X	X		43
61b			X		X	
61c	X		X			
61d		X	X			
62a		X	X			44
62b		X		X		

Appendix A (continued)

Student	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Number
62c		X			X	
62d	X	X				
64a	X	X				45
64b	X		X			
64c	X			X		
64d	X				X	
66a	X	X				46
66b	X		X			
66c	X			X		
66d	X				X	
67a		X	X			47
67b		X		X		
67c		X			X	
67d	X	X				
68a	X	X	X	X	X	48
68b	X	X	X	X	X	
68c	X	X	X	X	X	
68d	X	X	X	X	X	
69a				X	X	49
69b	X			X		
69c		X		X		
69d			X	X		
70a	X				X	50
70b		X			X	
70c			X		X	
70d				X	X	
73a	X	X	X	X	X	51
73b	X	X	X	X	X	
73c	X	X	X	X	X	
73d	X	X	X	X	X	
75a				X	X	52
75b			X		X	
75c		X			X	
75d	X				X	
77a			X	X		53
77b		X		X		
77c	X			X		
77d				X	X	
78a		X	X			54
78b	X		X			
78c			X		X	
78d			X	X		
79a	X	X				55
79b		X			X	

Appendix A (continued)

Student	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Number
79c		X		X		
79d		X	X			
80a	X				X	56
80b	X			X		
80c	X		X			
80d	X	X				
81a	X				X	57
81b	X			X		
81c	X		X			
81d	X	X				
82a	X	X				58
82b		X			X	
82c		X		X		
82d		X	X			
84a		X	X			59
84b	X		X			
84c			X		X	
84d			X	X		
85a	X	X				60
85b	X		X			
85c	X			X		
85d	X				X	
90a	X	X	X	X	X	61
90b	X	X	X	X	X	
90c	X	X	X	X	X	
90d	X	X	X	X	X	
91a				X	X	62
91b			X		X	
91c		X			X	
91d	X				X	
92a	X				X	63
92b		X			X	
92c			X		X	
92d				X	X	
93a	X	X				64
93b	X		X			
93c	X			X		
93d	X				X	
94a			X	X		65
94b			X		X	
94c	X		X			
94d		X	X			
95a		X	X			66
95b		X		X		

Appendix A (continued)

Student	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Number
95c		X			X	
95d	X	X				
100a	X	X				67
100b	X		X			
100c	X			X		
100d	X				X	
101a	X	X				68
101b	X		X			
101c	X			X		
101d	X				X	
102a		X	X			69
102b		X		X		
102c		X			X	
102d	X	X				
103a			X	X		70
103b			X		X	
103c	X		X			
103d		X	X			
104a	X	X	X	X	X	71
104b	X	X	X	X	X	
104c	X	X	X	X	X	
104d	X	X	X	X	X	
106a				X	X	72
106b	X			X		
106c		X		X		
106d			X	X		
107a	X				X	73
107b		X			X	
107c			X		X	
107d				X	X	
108a				X	X	74
108b			X		X	
108c		X			X	
108d	X				X	
109a			X	X		75
109b		X		X		
109c	X			X		
109d				X	X	
110a		X	X			76
110b	X		X			
110c			X		X	
110d			X	X		
112a	X	X				77
112b		X			X	

Appendix A (continued)

Student	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Number
112c		X		X		
112d		X	X			
114a	X				X	78
114b	X			X		
114c	X		X			
114d	X	X				
121a	X				X	79
121b	X			X		
121c	X		X			
121d	X	X				
129a	X	X				80
129b		X			X	
129c		X		X		
129d		X	X			
132a	X	X	X	X	X	81
132b	X	X	X	X	X	
132c	X	X	X	X	X	
132d	X	X	X	X	X	
133a	X	X	X	X	X	82
133b	X	X	X	X	X	
133c	X	X	X	X	X	
133d	X	X	X	X	X	
134a	X	X	X	X	X	83
134b	X	X	X	X	X	
134c	X	X	X	X	X	
134d	X	X	X	X	X	
136a				X	X	84
136b			X		X	
136c		X			X	
136d	X				X	
141a	X				X	85
141b		X			X	
141c			X		X	
141d				X	X	
142a				X	X	86
142b	X			X		
142c		X		X		
142d			X	X		
146a			X	X		87
146b			X		X	
146c	X		X			
146d		X	X			
147a		X	X			88
147b		X		X		

Appendix A (continued)

Student	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Number
147c		X			X	
147d	X	X				
148a	X	X				89
148b	X		X			
148c	X			X		
148d	X				X	
150a	X	X				90
150b	X		X			
150c	X			X		
150d	X				X	
152a	X	X	X	X	X	91
152b	X	X	X	X	X	
152c	X	X	X	X	X	
152d	X	X	X	X	X	
153a		X	X			92
153b		X		X		
153c		X			X	
153d	X	X				
155a	X	X				93
155b	X		X			
155c	X			X		
155d	X				X	
156a				X	X	94
156b	X			X		
156c		X		X		
156d			X	X		
157a	X				X	95
157b		X			X	
157c			X		X	
157d				X	X	
158a				X	X	96
158b			X		X	
158c		X			X	
158d	X				X	
160a			X	X		97
160b		X		X		
160c	X			X		
160d				X	X	
161a		X	X			98
161b	X		X			
161c			X		X	
161d			X	X		
164a	X	X				99
164b		X			X	

Appendix A (continued)

Student	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Number
164c		X		X		
164d		X	X			
166a	X	X	X	X	X	100
166b	X	X	X	X	X	
166c	X	X	X	X	X	
166d	X	X	X	X	X	
167a	X	X	X	X	X	101
167b	X	X	X	X	X	
167c	X	X	X	X	X	
167d	X	X	X	X	X	
169a	X	X				102
169b		X			X	
169c		X		X		
169d		X	X			
172a	X	X	X	X	X	103
172b	X	X	X	X	X	
172c	X	X	X	X	X	
172d	X	X	X	X	X	
174a			X	X		104
174b		X		X		
174c	X			X		
174d				X	X	
176a				X	X	105
176b			X		X	
176c		X			X	
176d	X				X	
177a	X				X	106
177b		X			X	
177c			X		X	
177d				X	X	
180a				X	X	107
180b	X			X		
180c		X		X		
180d			X	X		
181a			X	X		108
181b			X		X	
181c	X		X			
181d		X	X			
182a		X	X			109
182b		X		X		
182c		X			X	
182d	X	X				
184a	X	X	X	X	X	110
184b	X	X	X	X	X	

Appendix A (continued)

Student	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Number
184c	X	X	X	X	X	
184d	X	X	X	X	X	
186a	X	X				111
186b	X		X			
186c	X			X		
186d	X				X	
187a	X	X				112
187b	X		X			
187c	X			X		
187d	X				X	
188a		X	X			113
188b		X		X		
188c		X			X	
188d	X	X				
190a	X	X	X	X	X	114
190b	X	X	X	X	X	
190c	X	X	X	X	X	
190d	X	X	X	X	X	
191a				X	X	115
191b	X			X		
191c		X		X		
191d			X	X		
192a	X				X	116
192b		X			X	
192c			X		X	
192d				X	X	
193a				X	X	117
193b			X		X	
193c		X			X	
193d	X				X	
195a			X	X		118
195b		X		X		
195c	X			X		
195d				X	X	
198a		X	X			119
198b	X		X			
198c			X		X	
198d			X	X		
199a	X	X	X	X	X	120
199b	X	X	X	X	X	
199c	X	X	X	X	X	
199d	X	X	X	X	X	
201a	X	X				121
201b		X			X	

Appendix A (continued)

Student	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Number
201c		X		X		
201d		X	X			
203a	X				X	122
203b	X			X		
203c	X		X			
203d	X	X				
204a	X				X	123
204b	X			X		
204c	X		X			
204d	X	X				
205a	X	X				124
205b		X			X	
205c		X		X		
205d		X	X			
209a		X	X			125
209b	X		X			
209c			X		X	
209d			X	X		
211a			X	X		126
211b		X		X		
211c	X			X		
211d				X	X	
215a				X	X	127
215b			X		X	
215c		X			X	
215d	X				X	
217a	X				X	128
217b		X			X	
217c			X		X	
217d				X	X	
219a				X	X	129
219b	X			X		
219c		X		X		
219d			X	X		
221a	X	X	X	X	X	130
221b	X	X	X	X	X	
221c	X	X	X	X	X	
221d	X	X	X	X	X	
223a			X	X		131
223b			X		X	
223c	X		X			
223d		X	X			
224a		X	X			132
224b		X		X		

Appendix A (continued)

Student	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Number
224c		X			X	
224d	X	X				
225a	X	X				133
225b	X		X			
225c	X			X		
225d	X				X	
226a	X	X				134
226b	X		X			
226c	X			X		
226d	X				X	
228a	X	X				135
228b	X		X			
228c	X			X		
228d	X				X	

Note: a = RB, b = SB, c = JL, d = JS

X = rated

Appendix B

Behavioral Examples for Rating Dimensions

Introduction:

Scale	Rating	Definition
5	Excellent (could not be better	Introduction included First and Last name along with year in medical school and possible specialty
4	Very Good (little room for improvement)	Solid introduction, but less detailed than first and last name along with additional information. Included at least a hand shake
3	Good (Room for improvement, but a good solid student physician)	Introduction included only First name
2	Fair (Significant improvement is needed)	Introduced self while washing hands
1	Poor (Major weakness in this area)	Did not introduce self

Confidence:

Scale	Rating	Definition
5	Excellent (could not be better	Very decisive and unwavering during examination
4	Very Good (little room for improvement)	Good eye contact and decision making. Knocks and enters room immediately
3	Good (Room for improvement, but a good solid student physician)	Eye contact throughout exam. May have just gone through the motions
2	Fair (Significant improvement is needed)	Signs of nerves like tapping, clicking pen, tics. Knocks and hesitates to enter room.
1	Poor (Major weakness in this area)	Giggly or giddy. Completely unsure of his/herself

Appendix B (Continued)

Comfort:

Scale	Rating	Definition
5	Excellent (could not be better	Made the patient feel at ease with the diagnosis and potential procedure
4	Very Good (little room for improvement)	
3	Good (Room for improvement, but a good solid student physician)	Some signs of being unsure
2	Fair (Significant improvement is needed)	
1	Poor (Major weakness in this area)	Student completely unsure of him/herself and therefore unable to attend to patients needs

Listened:

Scale	Rating	Definition
5	Excellent (could not be better	Very attentive to patients questions. Clarified any concerns the patient had
4	Very Good (little room for improvement)	
3	Good (Room for improvement, but a good solid student physician)	Answered questions adequately
2	Fair (Significant improvement is needed)	
1	Poor (Major weakness in this area)	Did not listen to what the patient was saying. Ignored questions

Appendix B (Continued)

Consideration:

Scale	Rating	Definition
5	Excellent (could not be better	Student notes gestures and body language of SP and focuses on them until the student understands the meaning behind them (e.g., I notice you seem apprehensive, what can I do to help?)
4	Very Good (little room for improvement)	
3	Good (Room for improvement, but a good solid student physician)	Student takes into account gestures and body language of the SP and specifically comments on them (e.g., I notice you seem apprehensive)
2	Fair (Significant improvement is needed)	
1	Poor (Major weakness in this area)	Student ignores any gestures or body language from the SP

Terminology:

Scale	Rating	Definition
5	Excellent (could not be better	Broke down terms into easy to understand language. Answered all questions about medical jargon fully until the patient understood
4	Very Good (little room for improvement)	
3	Good (Room for improvement, but a good solid student physician)	Kept the terms simple or clarified when asked, but did not elaborate to make sure the patient fully understood
2	Fair (Significant improvement is needed)	
1	Poor (Major weakness in this area)	Continued to talk above the patients head even when asked for clarification

Conclusion:

Scale	Rating	Definition
5	Excellent (could not be better	Gave the patient a full debrief of what the current diagnosis was and how they would proceed
4	Very Good (little room for improvement)	
3	Good (Room for improvement, but a good solid student physician)	Let the patient know the next step, but did not give a synopsis of the what would be happening
2	Fair (Significant improvement is needed)	
1	Poor (Major weakness in this area)	No discernable conclusion

Cultural Competency:

Scale	Rating	Definition
5	Excellent (could not be better	Student is very aware of cultural impact of diagnosis and treatment (ex. If the patient has iron-deficiency anemia, the patient will need to eat more iron. One of the best sources of iron is from red-meat, but what if the patient is a vegetarian? For an excellent rating, the student would come up with a viable alternative)
4	Very Good (little room for improvement)	
3	Good (Room for improvement, but a good solid student physician)	Student is aware of any cultural impact the diagnosis and treatment will have on the patient
2	Fair (Significant improvement is needed)	
1	Poor (Major weakness in this area)	Student ignores any cultural aspects mentioned by the SP in the diagnosis and treatment

Appendix B (Continued)

Empathy:

Scale	Rating	Definition
5	Excellent (could not be better	Student truly feels what the patient is going through by stating things like “I understand how you feel”, “I know what you are going through”, etc.
4	Very Good (little room for improvement)	
3	Good (Room for improvement, but a good solid student physician)	Student feels sorry for the patient (sympathy), but does not have a true understanding of what the patient is going through by stating things like, “I know this must be hard for you”
2	Fair (Significant improvement is needed)	
1	Poor (Major weakness in this area)	Student is indifferent towards the patient

Partnering:

Scale	Rating	Definition
5	Excellent (could not be better	Student includes the patient in decision making process, making sure to double check with the patient before deciding on a treatment
4	Very Good (little room for improvement)	
3	Good (Room for improvement, but a good solid student physician)	Student includes the patient in decision making process
2	Fair (Significant improvement is needed)	
1	Poor (Major weakness in this area)	Student does not include the patient in the decision making process

Appendix B (Continued)

Honesty:

Scale	Rating	Definition
5	Excellent (could not be better	Student is open with the patient about their condition and possible treatments, even when faced with an emotionally reactive patient
4	Very Good (little room for improvement)	
3	Good (Room for improvement, but a good solid student physician)	Student is open with the patient about their condition and possible treatments
2	Fair (Significant improvement is needed)	
1	Poor (Major weakness in this area)	Student is not open with the patient by assuring them that everything will be alright no matter what (false reassurance)

Impact:

Scale	Rating	Definition
5	Excellent (could not be better	Student goes in depth asking about how diseases is impacting the patients life and family
4	Very Good (little room for improvement)	
3	Good (Room for improvement, but a good solid student physician)	Student briefly asks about how diseases is impacting the patients life and family
2	Fair (Significant improvement is needed)	
1	Poor (Major weakness in this area)	Student does not ask about how diseases is impacting the patients life and family

Appendix B (Continued)

Compassion:

Scale	Rating	Definition
5	Excellent (could not be better	Student is focusing 100% of their time and energy on the patient and shows additional compassion (ex. puts hand on patients shoulder)
4	Very Good (little room for improvement)	
3	Good (Room for improvement, but a good solid student physician)	Student is focusing 100% of their time and energy on the patient, but does not make any compassionate gestures (i.e. hand on the shoulder)
2	Fair (Significant improvement is needed)	
1	Poor (Major weakness in this area)	Student treats the patient more like a machine to be fixed than a person (i.e., no emotion in their voice or compassionate gestures)

Appendix C

Rater Correlations

In this section comprising Tables 1 thru 6, data are presented for how the raters correlated with themselves over the four different cases. The tables are broken down by dimension, CM, HX, and PX. All of these data come from the common item data for each dimension.

Table C1

Rater 1 Intraclass Correlations

CM				
	RB	*SB*	*JL*	*JS*
RB	1			
SB	0.244981	1		
JL	0.21833	0.025385	1	
JS	0.706387	0.062549	0.148764	1
HX				
RB	1			
SB	0.093433	1		
JL	0.306606	0.461073	1	
JS	0.034582	-0.14824	0.202692	1
PX				
RB	1			
SB	0.314712	1		
JL	-0.08801	-0.45238	1	
JS	0.186389	-0.10361	0.10114	1

Note: n = 20

Table C2

Rater 2 Intraclass Correlations

CM				
	RB	*SB*	*JL*	*JS*
RB	1			
SB	0.029828	1		
JL	-0.05764	-0.15159	1	
JS	0.072105	-0.1661	0.256386	1
HX				
RB	1			
SB	0.309118	1		
JL	-0.09141	0.566113	1	
JS	-0.25352	-0.23198	0.188877	1
PX				
RB	1			
SB	0.120513	1		
JL	-0.21544	-0.0909	1	
JS	0.189275	0.00766	0.077783	1

Note: n = 20

Table C3

Rater 3 Intraclass Correlations

CM				
	RB	*SB*	*JL*	*JS*
RB	1			
SB	0.368684	1		
JL	0.399168	0.415611	1	
JS	0.221987	0.554355	0.4998	1
HX				
RB	1			
SB	0.076084	1		
JL	0.503893	0.302626	1	
JS	0.051402	-0.02853	0.158533	1
PX				
RB	1			
SB	0.230851	1		
JL	-0.13829	-0.24628	1	
JS	0.034455	-0.08942	-0.24356	1

Note: n = 20

Table C4

Rater 4 Intraclass Correlations

CM				
	RB	*SB*	*JL*	*JS*
RB	1			
SB	0.150531	1		
JL	0.181585	0.313416	1	
JS	0.227636	0.054931	0.258652	1
HX				
RB	1			
SB	-0.26322	1		
JL	0.067065	0.284411	1	
JS	0.265631	-0.2622	-0.04157	1
PX				
RB	1			
SB	-0.02789	1		
JL	-0.18035	-0.34066	1	
JS	0.0517	-0.27423	0.180211	1

Note: n = 20

Table C5

Rater 5 Intraclass Correlations

CM				
	RB	*SB*	*JL*	*JS*
RB	1			
SB	0.396114	1		
JL	0.188706	-0.01746	1	
JS	0.243078	0.38763	-0.05944	1
HX				
RB	1			
SB	0.231666	1		
JL	0.500454	0.325163	1	
JS	0.452057	0.101368	0.263851	1
PX				
RB	1			
SB	0.50454	1		
JL	-0.28404	-0.28917	1	
JS	0.262709	0.232475	0.058621	1

Note: n = 20

Table C6

SP Intraclass Correlations

CM				
	RB	*SB*	*JL*	*JS*
RB	1			
SB	0.217989	1		
JL	-0.0988	-0.31779	1	
JS	0.550965	0.118381	0.195192	1
HX				
RB	1			
SB	-0.09266	1		
JL	0.230385	0.357806	1	
JS	0.357837	-0.14878	-0.13177	1
PX				
RB	1			
SB	0.419805	1		
JL	--	--	1	
JS	-0.03914	-0.15321	--	1

Note: n = 20

Appendix D

Individual Rater Descriptive Statistics

In this section, raters are broken up by case and then by dimension. These results are presented in Tables 1 thru Table 4. These scores were taken from the fully crossed example of 20 students so that the results would be directly comparable.

Table D1

Rachel Brown Fully Crossed Results by Rater

Dimension	Total Score	Mean	SD	Source
Overall	66	44.15	5.90	Rater 1
		43.0	6.09	Rater 2
		41.15	5.23	Rater 3
		44.4	5.48	Rater 4
		41.5	6.57	Rater 5
		55.15	6.35	SP
CM	35	25.60	2.93	Rater 1
		24.5	2.61	Rater 2
		23.85	2.48	Rater 3
		25.35	2.98	Rater 4
		24.15	3.20	Rater 5
		33.4	2.37	SP
HX	15	8.50	2.42	Rater 1
		8.2	2.73	Rater 2
		8.15	2.41	Rater 3
		8.50	2.76	Rater 4
		7.75	2.34	Rater 5

Dimension	Total Score	Mean	SD	Source
HX	15	9.85	3.01	SP
PX	16	10.05	3.02	Rater 1
		10.3	3.05	Rater 2
		9.15	2.50	Rater 3
		10.55	2.89	Rater 4
		9.60	3.25	Rater 5
		11.9	3.21	SP

Note: n = 20

Table D2

Samantha Browning Fully Crossed Results by Rater

Dimension	Total Score	Mean	SD	Source
Overall	91	67.3	7.35	Rater 1
		61	6.27	Rater 2
		63	4.38	Rater 3
		66.25	5.30	Rater 4
		61.95	5.91	Rater 5
		78.1	6.86	SP
CM	60	47.15	5.4	Rater 1
		41.75	3.42	Rater 2
		43.75	3.32	Rater 3
		45.5	3.98	Rater 4
		42.25	3.97	Rater 5
		54.4	5.11	SP

Appendix D (continued)

Dimension	Total Score	Mean	SD	Source
HX	15	9.8	2.33	Rater 1
		9.05	1.99	Rater 2
		9.55	2.39	Rater 3
		9.5	2.61	Rater 4
		8.9	2.38	Rater 5
		11.7	2.05	SP
PX	16	10.35	2.64	Rater 1
		10.2	2.84	Rater 2
		9.7	2.64	Rater 3
		11.25	2.45	Rater 4
		10.8	2.48	Rater 5
		12	2.73	SP

Note: n = 20

Table D3

John Long Fully Crossed Results by Rater

Dimension	Total Score	Mean	SD	Source
Overall	75	52	6.49	Rater 1
		44.4	3.90	Rater 2
		49.65	4.18	Rater 3
		50.9	4.91	Rater 4
		45.6	5.25	Rater 5
		66.8	6.85	SP

Appendix D (continued)

Dimension	Total Score	Mean	SD	Source
CM	60	43.6	5.33	Rater 1
		36.45	2.93	Rater 2
		41.1	3.55	Rater 3
		42.55	3.99	Rater 4
		37.3	4.38	Rater 5
		54.15	5.83	SP
HX	11	7.4	1.43	Rater 1
		7.15	1.60	Rater 2
		7.5	1.36	Rater 3
		7.4	1.14	Rater 4
		7	1.26	Rater 5
		9.35	1.79	SP
PX	4	1	1.17	Rater 1
		.8	1.06	Rater 2
		1.05	1.15	Rater 3
		.95	1.19	Rater 4
		1.3	1.26	Rater 5
		3.3	1.8	SP

Table D4

John Sexton Fully Crossed Results by Rater

Dimension	Total Score	Mean	SD	Source
Overall	56	36.05	5.26	Rater 1
		33.85	4.75	Rater 2

Appendix D (continued)

Dimension	Total Score	Mean	SD	Source
Overall	56	33.9	5.47	Rater 3
		35.15	5.41	Rater 4
		30.35	5.87	Rater 5
		45.05	9.63	SP
CM	35	24.75	3.19	Rater 1
		23.35	2.25	Rater 2
		22.40	3.79	Rater 3
		24.25	3.68	Rater 4
		20.9	3.37	Rater 5
		30.05	6.61	SP
HX	12	6.1	2.10	Rater 1
		5.45	2.46	Rater 2
		5.9	2.38	Rater 3
		5.95	2.11	Rater 4
		5.2	2.17	Rater 5
		5.95	2.11	SP
PX	9	5.2	1.79	Rater 1
		5.05	2.06	Rater 2
		5.6	1.85	Rater 3
		4.95	1.67	Rater 4
		4.25	2.31	Rater 5
		4.95	1.67	SP

About the Author

Frederick R. B. Stilson was born on October 7, 1980 in Atlanta, Georgia. He graduated from George Walton Comprehensive High School in Marietta, Georgia in 1999. He received his Bachelor of Science in Psychology from the University of Georgia in 2003. He then moved to Tampa, Florida to attend graduate school at the University of South Florida. He received his Master of Arts degree in Industrial/Organizational Psychology in 2006.